Computational and Perceptual Characterization of Auditory Attention

by

Emine Merve Kaya

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August, 2017

Abstract

Humans are remarkably capable at making sense of a busy acoustic environment in real-time, despite the constant cacophony of sounds reaching our ears. Attention is a key component of the system that parses sensory input, allocating limited neural resources to elements with highest informational value to drive cognition and behavior. The focus of this thesis is the perceptual, neural, and computational characterization of auditory attention.

Pioneering studies exploring human attention to natural scenes came from the visual domain, spawning a number of hypotheses on how attention operates among the visual pathway, as well as a considerable amount of computational work that attempt to model human perception. Comparatively, our understanding of auditory attention is yet very elementary, particularly pertaining to attention automatically drawn to salient sounds in the environment, such as a loud explosion. In this work, we explore how human perception is affected by the saliency of sound, characterized across a variety of acoustic features, such as pitch, loudness, and timbre. Insight from psychoacoustical data is complemented with neural measures of attention recorded

ABSTRACT

directly from the brain using electroencephalography (EEG). A computational model of attention is presented, tracking the statistical regularities of incoming sound among a high-dimensional feature space to build predictions of future feature values. The model determines salient time points that will attract attention by comparing its predictions to the observed sound features. The high degree of agreement between the model and human experimental data suggests predictive coding as a potential mechanism of attention in the auditory pathway.

We investigate different modes of volitional attention to natural acoustic scenes with a "cocktail-party" simulation. We argue that the auditory system can direct attention in at least three unique ways (globally, based on features, and based on objects) and that perception can be altered depending on how attention is deployed. Further, we illustrate how the saliency of sound affects the various modes of attention.

The results of this work improve our understanding of auditory attention, highlighting the temporally evolving nature of sound as a significant distinction between audition and vision, with a focus on using natural scenes that engage the full capability of human attention.

Primary Reader: Mounya Elhilali

Secondary Reader: Ralph Etienne-Cummings

Acknowledgments

I would first like to thank my advisor, Mounya Elhilali, for her friendly guidance, patience, and support throughout my PhD work. She is a true role model both as a scientist and as an individual. I would like to thank Professor Rene Vidal for his mentorship in my first few semesters at Johns Hopkins. I am also grateful to Professors Ralph Etienne-Cummings, Ernst Niebur, and Andreas Andreou for being part of my dissertation committees, and for providing valuable insight to my research along the way.

I am indebted to my friends in the Laboratory of Computational Audio Perception for making our time in the lab an enjoyable experience. Mike Carlin, Sridhar, Kailash, Dimitra, Nick, Susan, Mike Wolmetz, Debmalya, Ashwin, Ben: It was a pleasure to have all of you as part of my PhD experience. Thank you for the long conversations, conference adventures, patience with my coffee machines, and letting me keep up Christmas decorations until February. Special thanks to Nick, for his endless generosity and always going above and beyond whenever I needed help. Many thanks to my friends in the Vision, Dynamics, and Learning Lab: Jixin, Annalisa,

ACKNOWLEDGMENTS

Huong, thank you for your friendship and the crab cakes. Avinash, Dheeraj, Ertan, Ehsan, Rizwan, Roberto, many thanks for your kind and patient support in so many ways. I owe Ertan and Muge so much for their welcoming friendship when I first came to Baltimore, and treating me like family. Thanks are also due to countless friends in the GRO and ECE with whom we went through many interesting ordeals, classes and experiences.

I would like to thank my family for their unconditional love, believing in me, and expanding my horizons. Last, but not least, thank you, Noyan, for being a great partner in endless academic, life and travel adventures.

Dedication

To my family, for their love and support

Contents

List of Tables

List of Figures

LIST OF FIGURES

Chapter 1

Introduction

Humans have a remarkable ability to make sense of complex acoustic information that is constantly reaching our ears. The auditory system can segment, recognize, and localize sources of sound with little to no effort on our part. This system is flexible: Under highly noisy conditions, we can focus our attention to improve perception of certain sound sources, but we would still notice when someone new begins talking in a meeting, and we turn around to look for who mentioned our name. The auditory system automatically evaluates important information for us (new speakers, our name) but allows us to ignore distracting sounds to focus on a single source as well. These dynamics are part of auditory scene analysis, a proposal of how sound signals that enter through our ears all together are separated and processed to drive perception.

Attention is a key processing mechanism that has significant effects on how sen-

sory information translates to meaning and action. On a busy street, surrounded by vehicles, many conversing groups passing by, alarm noises, music, animals, we are not overwhelmed by the influx of sound, because we do not process it all to the same extent. Attention is thought to be similar to a filtering mechanism that allows some sound sources to progress to higher areas in a hierarchy of information processing. There are many factors that determine which sounds will benefit from detailed analysis, among them our goals, expectations, memory, alertness, and the inherent salience of external sounds (such as a loud explosion). In the same street, different people might not all perceive the same sounds, even though the same signals are reaching both of their ears.

Because the attending human auditory system is so efficient (performing all the complex analysis and interpretation steps in real-time) and robust to noise, it can serve as a valuable reference for developing acoustic processing technologies that can extract key information from large data given different goals. With the amount of sound data increasing in great amounts every year, it is becoming more and more necessary to create efficient computational models that can make sense of unstructured natural audio. Understanding and being able to program insights from the human attention system would have significant implications in artificial intelligence technologies rapidly becoming a reality, from having a secretary in your pocket to household robots that react to commands. Such systems must be able to filter the commands from irrelevant background sounds for interpretation, precisely the task

performed every day by human auditory attention.

1.1 Motivation

Although attention is a major component of perception, we have merely an elementary understanding of how the brain accomplishes this complex feat. Further, most of the research and theories of human attention have come from the visual domain, with little thought given to other sensory modalities. That being said, extensive research in audition has characterized the organization of the auditory pathway, basic aspects of neural coding mechanisms that represent sound in the brain, and psychoacoustical phenomena that contribute to auditory perception. With respect to attention, much of the existing work focuses on describing how directed attention affects the neural responses or perceptual factors, for example, the effect of attention on formation of auditory objects.

One of the primary components of attention that has a very sparse representation in the literature is saliency driven bottom-up attention (eg. attentional direction to a loud explosion). The few studies that investigate auditory saliency mostly follow the lead of theories from vision, where this topic has been thoroughly explored. In vision, salience is very intuitive, a red bird in a green forest is salient. Importantly, confronted with a picture of the aforementioned forest, our eyes will immediately look at the bird first. In general, without any goal or task, humans tend to analyze

a visual scene in order of the most to least salient sections, and this visual search can be recorded with eye-tracking devices. Such recordings give us a ground-truth for what humans find salient, the types of stimuli that will attract their attention in the absence of a task. This paradigm provides a solid foundation to investigate further forms of attention: What level of salience is distracting, how does saliency interact with goal-directed attention, what are the processing stages associated with the selection of objects that will make it to the forefront of consciousness, among others.

The lack of a direct parallel to eye-tracking of images in audition has been one of the biggest reasons that so little is yet known about auditory saliency. Although we can consider auditory saliency intuitively as well, such as a loud sound being salient, a significant difference in modalities limits the generalizations that can be made. In the case of vision, a whole scene is presented to our eyes at once and remains static while it is analyzed. In audition, the whole scene evolves over time. Specifically, the problems involved with studying auditory saliency are:

1. How can we record auditory saliency? While behavioral experiments can be conducted to probe perception of salience, it is challenging to create a design that will eliminate confounding cognitive factors, including goal-directed attention.

2. How can we define auditory saliency? In vision, contrast in visual features is the most basic cause of salience, such as color, orientation, or intensity. What types of difference among acoustic features are perceived as salient?

Beyond saliency driven attention, decades of research has established a baseline for identifying key processes of acoustic scene analysis, yet little emphasis has been put on the effect of attention in parsing the auditory environment. Meanwhile, visual research has determined that attention can be directed to different components of a scene, based on location, feature (such as searching for a red book in a bookcase), or object (such as searching for a book on a desk), although it is still uncertain how these forms of visual attention come together to drive perception. Although recent behavioral and neural imaging results have started to demonstrate that attention can operate on auditory objects, much is unknown about whether auditory attention can be deployed to different components of the scene, how these different forms of attention are related to each other, and to what extent different sensory modalities share common attention mechanisms.

1.2 Goal and Approach

The overall goal of this dissertation is to characterize auditory attention in natural-sounding scenes. To that end, the first step is to derive a paradigm to obtain unbiased perceptual data from humans to establish a baseline for how the outside world shapes our attention without active control. Multiple approaches to obtaining ground-truth auditory saliency data are tested. Behavioral measures are valuable for their ease of implementation and interpretation, however, the extent to which volitional factors

are reflected in behavioral data is a concern. Due to this reason, neural measures of bottom-up attention are examined. Although it is challenging to interpret neural responses to complex sounds, we extract various markers of bottom-up attention processing directly from the brain.

To characterize the process of sounds attracting attention, a computational model is developed, testing the hypothesis that bottom-up attention is driven by violation of predictions among acoustic features. The model builds on biologically-inspired principles to match human perception, providing a possible explanation for attention processing among the auditory pathway.

Next, we explore how humans actively attend to sound, and how this top-down, directed attention is affected by the inherent salience of events in the auditory environment. We extend the previous experimental paradigm with attention directed to various components of the scene, allowing the characterization of bottom-up and top-down attention in a naturalistic setup.

1.3 Contributions

In addressing the challenges and goals outlined above, the main contributions of this dissertation can be summarized as follows.

1. A fundamental contribution is the treatment of auditory attention in a domain specific manner. Previous studies have commonly tried to interpret acoustic

scenes as a two-dimensional image, with time and frequency as axes instead of visual space. Once this transformation is achieved, the scene can be analyzed similar to visual principles. Treating the time dimension as a static feature discards important information in audio. Sound is a temporal entity. A more appropriate approach to analyze auditory information is to consider acoustic features evolving over time, and the learning and adaptation that occurs as a result.

2. A behavioral methodology to probe bottom-up auditory attention is developed. This method examines the perception of saliency systematically among a comprehensive set of acoustic features. Importantly, the features evaluated can covary, allowing the investigation of feature interactions in auditory perception. This methodology also leads to a possible definition of saliency as a deviation from the ongoing acoustic regularities in a scene.

3. A computational attention model is presented that extracts a rich acoustic feature space and models statistical predictions based on sound that has been heard so far. It extracts salience information based on how well the incoming sound matches predictions. The model is demonstrated to predict human responses.

4. Using a variety of signal processing techniques, automatic attention markers are extracted from neural electroencephalogram (EEG) data. Traditional processing of EEG data requires well-separated short sound segments (most com-

monly pure tones) repeated many times to detect neural responses to stimulus. Although completely natural scenes have not yet been tested, this approach provides a foundation to probe attention to short time segments in complex natural soundscapes from EEG recordings.

5. The effect of different types of volitional attention are contrasted for the same acoustic scenes, demonstrating that perception can depend on how attention is deployed, and that directed-attention interacts with saliency.

1.4 Overview

This dissertation is divided into four chapters. Chapter 2 provides a detailed background to the auditory attention literature from a modeling perspective. Top-down and bottom-up attention are introduced, as well as the current state-of-the-art in modeling both forms of attention. This chapter also considers the challenges associated with recording, defining, and modeling auditory attention, along with how previous studies have tried to handle these problems. Practical avenues where attention research can make an impact, and descriptions of a variety of computational systems that have made use of attention mechanisms are presented.

In Chapter 3, bottom-up auditory attention is defined in a system building statistics among a high-dimensional feature space. A computational model is developed, testing the hypothesis of predictive coding as a mechanism underlying auditory at-

tention. To test this hypothesis, human perceptual data is collected, with results contrasted with predictions from the computational model. The computational model is able to give close matches to human performance, even without using human data for training. The importance of considering feature interactions is highlighted, a crucial component resulting in the success of the model. This chapter also discusses the significance of considering time in auditory attention, and explains the temporal build-up of attention.

Chapter 4 attempts to record an unbiased measure of bottom-up attention by exploring neural responses obtained by EEG from unattending subjects hearing the same stimulus that was used to collect behavioral responses in Chapter 3. The complexity of the scenes, designed to sound natural-like for the behavioral experiment, presents a number of processing challenges. Despite the decreased signal-to-noise ratio, we extract a variety of markers that correlate with levels of salience. Importantly, these markers do not only reflect the acoustics of the sounds being heard, but represent attention being drawn in an exogenous manner.

We look at how different forms of engaging attention influence perception in Chapter 5. Attention is directed towards objects, or different features, and contrasted with exogenous attentional draw in free-listening. The emphasis on natural stimuli is preserved, as is the notion of saliency as a deviation from the acoustic regularity of the scene. This chapter demonstrates that attention can operate in three unique mechanisms, and that directed attention is influenced by the inherent salience of sounds.

Finally, Chapter 6 summarizes the key results from this dissertation, and discusses avenues of further research.

Chapter 2

A review of auditory attention from a modeling perspective

2.1 Introduction

While at a cocktail party, we often find ourselves flooded by a cacophony of sounds that impinge on our ears from a multitude of sources. The challenge of directing our attention despite numerous prominent distractors, referred to as the "cocktail party problem" [1, 2], engages intricate neural networks and cognitive processes that enable the brain to parse information in the environment [3]. These processes allow us to navigate our surroundings, focus on conversations of interest, enjoy the background music, and be alert to any salient sound events such as someone calling our name or the ringtone of our phone. Throughout this scene analysis process, attention plays a

crucial role in mediating both perception and behavior by focusing both sensory and cognitive resources on pertinent information in the stimulus space [4]. This article provides a review of modeling studies of auditory attention and their impact on studies of attention in audition.

Attention is not a single, unidirectional process [5, 6]. It can be modulated by "bottom-up" stimulus-driven factors, as well as "top-down" task-specific goals, expectations, and learned schemas. Ultimately, attention is a form of information bottleneck that samples the massive sensory input constantly impinging on our ears and directs sensory and cognitive resources to the most relevant events in the soundscape [7]. Owing to the complexity of auditory scenes, the relevance of a sound event can be dictated by the scene itself (e.g. a conspicuous sound event such as a gunshot that would attract attention) or by a task at hand (e.g. to follow a conversation with a friend amidst competing sound sources).

While attention has started to garner increasing interest from the auditory research community [8, 9, 10, 11], there is not much tradition of developing computational models of attention in the context of sound systems. Such models would need to account for the auditory system's ability to adapt to the demands of an ever-changing acoustic environment and task goals. Recent physiological findings have been amending our views of processing in the auditory system, replacing the conventional view of static processing in sensory cortex with a more "active" and malleable mapping that rapidly adapts to the tasks at hand, sound context, and listening conditions [11]. Nu-

merous studies have revealed that our auditory experiences can have significant local effects by transforming receptive field properties of individual neurons, and profound global effects by reshaping cortical circuits [12, 13]. These effects extend beyond early sensory areas and indicate attentional modulation throughout the auditory cortex, shedding light on the distributed nature of processing in auditory pathways in the context of cocktail party settings [14].

Although research on the neural underpinnings of these networks is thriving, our understanding of the exact role of adaptive stimulus- or task-directed processing remains in its infancy. The field is particularly challenged by the lack of theories that integrate our knowledge of cortical circuitry in the auditory pathway with adaptive and cognitive processes that shape behavior and perception of complex acoustic scenes. In contrast, active and adaptive processing has more commonly been explored in models of the visual system. These implementations typically model predictive coding in the visual thalamus (LGN), contextual modulation in primary visual cortex (V1), attentional modulation in higher cortical areas (V2 and V4, area MT), and decision making in parietal and frontal cortex [15, 16]. That being said, recent theoretical studies are providing insight into common processing traits of active attention across modalities [17].

A number of perspectives have emerged about conceptual frameworks for understanding the role of attention in auditory perception. Much of this work closely parallels theories from vision in which attention is viewed as a multifaceted phe-

nomenon that encompasses mechanisms of selection, integration and sampling [18]. In one view, attention can be considered as a filtering or a selection mechanism. This interpretation ties in directly with findings of receptive field properties in sensory cortex, whereby neurons can be viewed as filters whose properties are modulated by task-directed attention and whose activity can be adapted depending on sensory contexts [11, 19]. At a larger scale, this view extends to object-based or semantic selection processes whereby attention to a specific target or class of sounds (e.g. speech, music) would engage specific neural circuits [14, 20]. This view parallels selection theories in vision, which present frameworks for funneling only relevant information to the processing pipeline, either at an early or later stage, acting as an informational bottleneck that mitigates the limited computational resources of the sensory system [5]. An alternative view of attention frames it as an integration mechanism, whereby attentional feedback acts as a prior to bias processing of certain stimuli of interest. Many theories of sound perception in complex settings favor this view, under which attention operates as a "glue" that binds together elements belonging to the same event. This interaction between object formation and selective attention is instrumental in guiding the organization of the foreground and background, and the interaction between the perceptual representations of sound targets and interferers [21, 22].

The present review aims to provide a synopsis of current computational efforts in modeling attention in the context of auditory scene analysis. Fig. 2.1 provides a general overview of models included in this review. These models often cluster

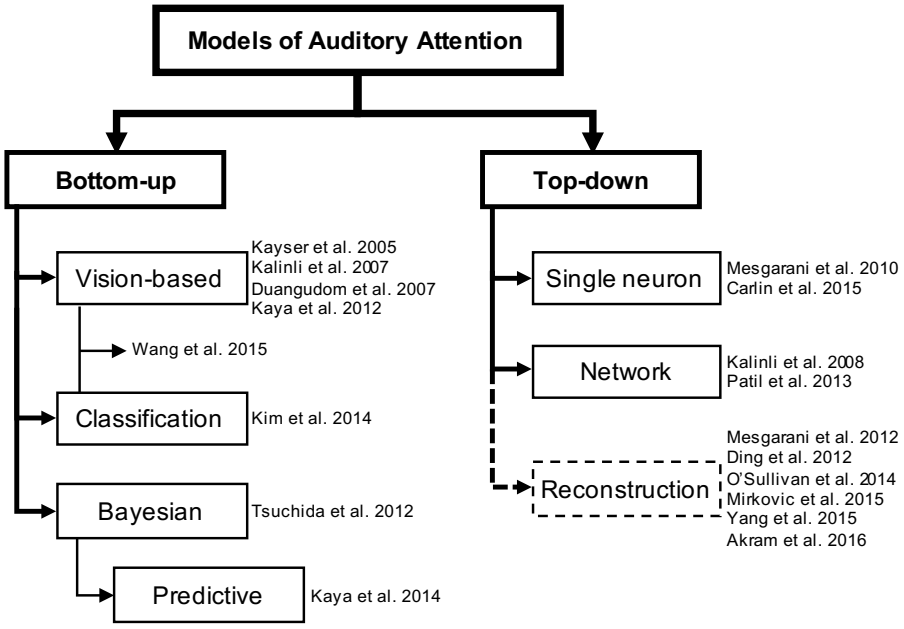

Figure 2.1: A broad classification of auditory attention models described in this chapter. Reconstruction techniques are not computational models in the traditional forward architecture of sound to "perception"; however, this methodology provides valuable insight in understanding task-directed attention.

around accounts of bottom-up or top-down processing, though they remain confined by handpicked experimental observations. The article reviews the relevant perspectives for both sensory-driven and task-driven attentional models and discusses some efforts to validate such models. The review also touches on relevant applications of such models in audio systems and hearing technologies.

2.2 Models of auditory attention

2.2.1 Bottom-up attention

Models of bottom-up attention remain very scarce in the auditory literature. The limited efforts in this direction have greatly benefited from the very prolific research on bottom-up attention (or salience) in vision. Indeed, visual salience is a thriving research field that has resulted in a rich body of work examining the perceptual attributes underlying visual salience [23], as well as its behavioral correlates [24] and underlying neuroanatomy [25, 26]. In parallel, computational models of visual salience have built on this knowledge and made use of the availability standardized eye-tracking datasets to develop detailed Bayesian and hierarchical accounts of visual perception [27]. These models can not only account for human behavior in natural scenes, but are able to expand the possibilities of computer vision applications to tackle challenging visual scenes in fields such as robotics, medical imaging and surveillance systems [28, 29, 30].

Building on this tradition in the visual modality, early models of auditory bottom-up attention adapted popular visual salience models to the domain of sound. Kayser et al. presented one of the early efforts in this direction [31]. This work treated the time-frequency representation of sound as an "auditory image" from which spectro-temporal features such as intensity and spectro-temporal contrast can be extracted to parallel the feature analysis process in vision models. The back end of the model was

essentially a visual salience model, in which all features were scaled to generate multi-scale maps which were then normalized to highlight conspicuous peaks and integrated to provide an auditory salience map. Though operating on relatively simple features and adopting a vision-based integration architecture, this model was able to reliably match both human and monkey behavioral responses in tasks involving detection of salient sounds embedded in different backgrounds. This work not only demonstrated that salience processing in the brain may share commonalities across sensory modalities, but it also provided a guide to designing psychoacoustical experiments to probe auditory bottom-up attention in humans.

This initial effort was later expanded to incorporate more intricate analyses of auditory features. Work by Kalinli et al. [32] operated on the same auditory image and salience extraction architecture but extended the feature set to include pitch and orientation along both time and frequency, hence incorporating more relevant auditory cues. It also provided an improved contrast computation scheme to derive feature maps, making them more robust to noise and multiple salient locations. Duangudom et al. [33] extended the feature analysis to incorporate more biologically plausible mechanisms that mimic processing in the peripheral and central auditory system [34]. This analysis allowed the derivation of spectro-temporal modulation features that simulate neural responses in the mammalian auditory cortex. These neural-like processes provided a multi-scale mapping of the incoming auditory stimulus, effectively replacing the parallel feature maps favored in earlier auditory salience models.

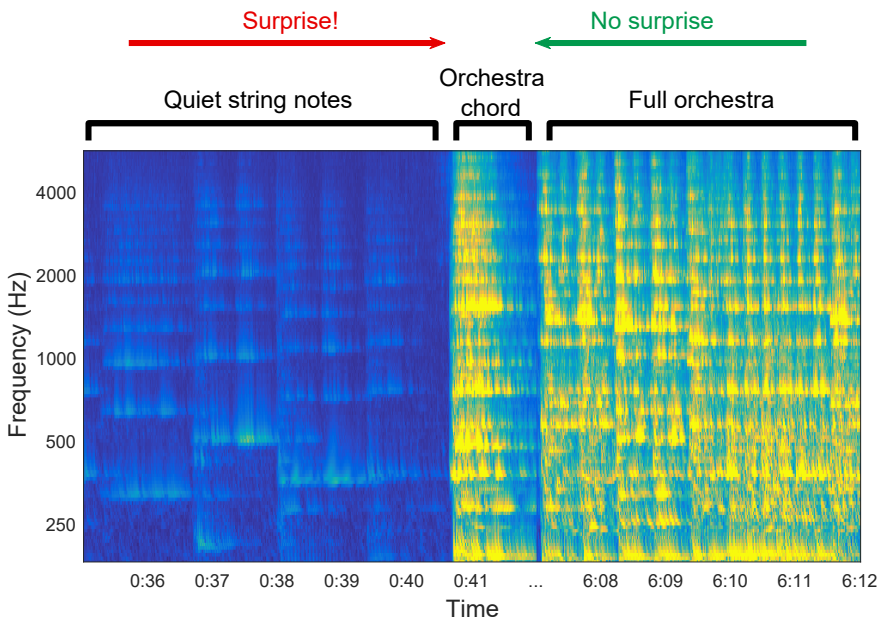

Figure 2.2: The spectrogram (time-frequency "image") of an excerpt from Haydns Surprise Symphony. Marked times correspond to the approximate location in the second movement. The surprising section is a loud chord played by the entire orchestra following a long passage of quiet string instruments. We consider the scenario of an orchestral passage immediately following the surprise chord. If the passage were reversed in time, the surprise chord would no longer be surprising, and the switch to a quiet passage is not as surprising as the switch to a sudden loud passage. This figure demonstrates the dependence of auditory salience on time and context.

While the salience analysis was similar in essence to that for vision-based models, this study began to steer the literature towards placing an emphasis on biological plausibility.

Despite their relative success in extending vision-based frameworks to audition, all of the aforementioned models failed to account for an important distinction between auditory and visual processing, notably the the nature of sound as a temporally evolving entity. By treating the time(T)-frequency(F) spectrogram as an auditory image, these models treated the T-F dimensions as spatial X-Y axes, failing to process

the time axis as a special dimension. Effectively, the "auditory image" approach ignores effects of temporal buildup and short and long-term dependencies, and results in non-causal analyses of current events based on future information. Consider for instance a musical scene such as Haydn's "Surprise" symphony (Fig 2.2): A mellow string passage abruptly interrupted by a loud, full orchestra chord a highly salient section. If the chord was repeated shortly after, you might be surprised again, but not as much as the first time, as you have now adjusted your expectations as to what might occur in the piece. If this chord were to start regularly repeating, it would eventually blend into the music and attract little attention. Now consider if this scene were played backwards, so the loud chord was heard repeatedly from the onset. None of the occurrences would surprise the listener the salience has disappeared. The surprise only works when the music is considered as a temporal entity.

One of the first models to address this problem computed a temporal salience map similarly to the model of Kayser et al., but considered all of the features as evolving temporally, rather than as two-dimensional images [35]. The feature space was expanded to include perceptual properties of sound: loudness, pitch and timbre. All features were analysed over time to highlight their dynamic quality before normalizing and integrating across feature maps in line with vision-based models. By contrast, Tsuchida & Cottrell [36] adapted a different, statistics-based approach from the vision literature [37]. Their implementation combined long-term scene statistics computed from natural sound samples with local, rapidly changing statistics of the

current incoming sound. In this framework, salience became a probabilistic account, where a sound is flagged as salient if it is determined to be unusual relative to learned statistics. This model was also the first to consider the computational efficiency of the features used, where a cochleogram was adopted instead of a spectrogram and principal component analysis was applied to reduce feature dimensionality while retaining significant variations in the features.

Even with the advances achieved by temporal salience models, basing attentional mechanisms on processes from the visual domain inherently limits the capabilities of an auditory salience model. Recognizing this, efforts in modeling auditory attention began shifting from adaptations of the visual literature to building upon inspiration from mechanisms known or hypothesized to take place in the auditory pathway. As this research area is yet in its early stages, there is an array of possible mechanisms to explore, and the following models have explored different avenues to modeling bottom-up auditory attention.

Kaya & Elhilali [38] proposed the first auditory attention model that was not based on a vision equivalent, but was rather motivated by processing known to occur in the auditory pathway. This model explored the role of predictive coding and theories of auditory deviance detection as possible underlying mechanisms determining auditory salience in the brain [39, 40, 41]. This approach puts great emphasis on the role of processing events over time and shaping neural responses of current sounds based on their preceding context. Kaya & Elhilali employed a rich feature

space modelling human perception of sound [34]. This model mapped the acoustic waveform onto a high-dimensional auditory space that explicitly encoded perceptual loudness, pitch and timbre of the incoming sound, building upon evolving temporal features [38]. The attention model collected feature statistics over time and made predictions about future sensory inputs. Salient times were flagged as those for which the incoming features differed significantly from expectations. Another novel aspect of this model was the role of integration across features in guiding salience predictions. Earlier models typically employed a simple linear combination across features with a fixed weight for each feature. The Kaya & Elhilali model rejected the notion of independence across auditory features of a complex scene in guiding salience perception. Instead, the model proposed a nonlinear interaction across the feature space, implemented by asymmetrical weights between pairwise features, and guided by psychoacoustic experiments.

Two trends emerged from this work that are reflected in most current auditory attention models: building probabilistic expectations of sound to derive salience, and employing behavioral responses from perceptual experiments with human listeners to learn properties of acoustic features relevant for salience perception. The idea that salience is derived from statistics gathered over the scene was further explored in the work of Wang et al. [42]. This study computed Shannon entropy as a measure of the informational value of incoming sound segments, and classified them as salient or ordinary depending on whether they contained a high amount of information.

This is in line with the concept that bottom-up attention alerts us to important events in a scene. Moreover, the study by Wang et al. offered a composite system that combined parallel paths including: (i) a temporal analysis of sound features operating on different components derived from mel-frequency cepstral coefficients [43], an alternative and very popular way to represent frequency features based on perceptual measures of pitch; (ii) a spectral mapping analysing the power spectral density of the stimulus and (iii), the image salience model based on mechanisms by Kayser et al. [44]. This composite system demonstrated the benefits of extending the vision-based model and provided further robustness to salience estimates especially in real noisy soundscapes.

In contrast with more theoretical approaches to auditory salience, Kim et al. [45] took a more data-driven approach by employing human behavioral judgements of salience to train a linear classifier that performed a simple filtering followed by feature integration based on data-driven weights. Behavioral data were gathered by subjects annotating salient locations in natural recordings of conference room meetings, and these data were used to train a model that maximized the separation between the salient and non-salient sound segments in the feature space. The results revealed that the emerging discriminant was shaped to detect temporal and frequency contrasts, and most specifically worked as an onset detector. Tordini et al. [46, 47] approached the problem from the opposite direction: while Kim et al. used no prior knowledge of acoustical features to guide their feature estimation, Tordini et al. tested

the contribution of acoustic features in defining auditory salience. Features such as temporal centroid, spectral centroid, harmonicity, effective duration and tempo were all found to correlate with salience ratings. The results also revealed interactions between some of these features in line with observations from Kaya & Elhilali [38].

It is worth highlighting that one of the challenges of studies of auditory salience is the open interpretation of what auditory salience refers to. Visual salience has historically relied on measures of eye gaze despite their shortcomings [28, 48, 49]. In audition, the lack of unified metrics to define salience remains a major challenge. Unmistakably salient scenarios such as a loud explosion or a male talking amongst females result in large enough loudness or pitch differences that every auditory salience model should be able to detect outlier events. However, more intricate processing is necessary for auditory events that are not as objectively salient, such as noticing a cricket among cicadas. The simple image-based features extracted in most of the aforementioned models are insufficient to capture subtle changes in temporal dynamics. Furthermore, feature interactions play an important role in determining perceived salience [38, 47] a factor unaccounted for in most models.

2.2.2 Top-down attention

In contrast with bottom-up attention, top-down models of auditory selective attention build on a richer body of work investigating the neural underpinnings of task-driven attention in the auditory system. It is well established that neural activ-

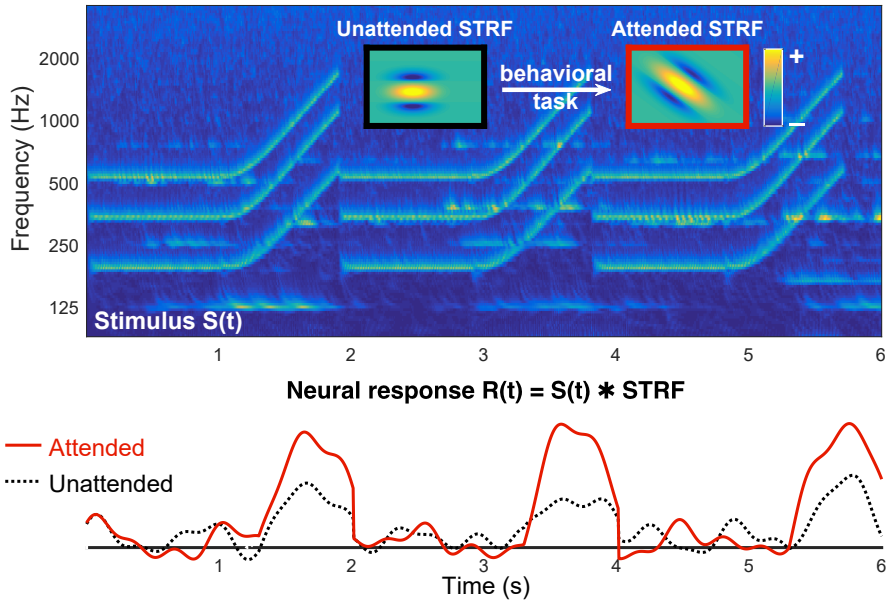

Figure 2.3: Attending to a particular sound characteristic tunes the neural spectro-temporal receptive fields (STRFs) and boosts the neural signal at times of attended event. Violin notes are overlaid with frequency modulations (FMs), illustrated with the spectrogram S(t). When instructed to attend to the FM segments, the STRF adapts to the orientation of the modulations, resulting in an enhancement in the neural response R(t).

ity across the auditory cortex is heavily modulated by directed attention [9, 13, 50]. Hubel et al.'s early findings in the late 1950s [51] showed modulation of neural activity of single units in cat auditory cortex when animals paid attention to novel or surprising acoustic events, such as jingling of keys. They dubbed such neurons "attention units" in the auditory cortex. Since then, many studies have reported similar "attention" effects under controlled behavioral conditions, in different animal models and across various auditory cortex regions.

Characterization of the tuning properties of cortical neurons using computational techniques has played a major role in investigating adaptive effects of attention on

cortical activity. Specifically, spectro-temporal receptive fields (STRFs) are mathematical descriptions of the selectivity of individual neurons in response to sound events [52]. The STRF is a two-dimensional timefrequency representation of the tuning properties of cortical neurons (Fig 2.3). From a systems theory viewpoint, each neuron can be thought of as a filter whose STRF describes the timefrequency attributes that excite the neuron [53, 54]. Evidence from behaving animals revealed that as behavioral goals changed, the tuning characteristics of individual neurons as captured by their STRFs adapted rapidly [55, 56, 57]. This neural adaptation, or rapid plasticity, plays a role in enhancing neural responses to temporal and spectral modulations belonging to the target sound events, the foreground, and suppressing those that fall outside the target, the background (Figure 2.3). Effectively, under control of attention, the neural population appears to increase the contrast between the target and background, hence facilitating focusing on sound events of interest [11]. Crucially, this process appears to be rapid, induced by attention, dependent on task and reward structure. It reflects the behavioral state of the animal [58] and spans both primary and higher auditory areas [59].

Beyond findings at the single-neuron level in animal models, various non-invasive techniques have been used to investigate the extent of attentional modulation across auditory cortex for more complex auditory scenes in human listeners. Results using functional magnetic resonance imaging and electroencephalogram (EEG) have confirmed attention-driven increase of neural activity in the auditory cortex [60, 61].

Neural effects revealing distributed activity induced by spatial and non-spatial forms of auditory attention have also been observed [14, 62]. Different types of attention, notably feature-based versus object-based attention, appear to induce differential activation engaging areas such as planum temporale and different regions of the superior temporal gyrus [63, 64]. Tying back to results from single units in animal models, recent advances in computational methods allowed analysis of neural recordings in human listeners using magnetoencephalography (MEG) and surface electrodes that revealed greater activation to attended sounds relative to unattended sounds [9, 65]. Going further, mathematical tools are now being developed to allow estimation of ensemble receptive fields from MEG and EEG recordings, laying promising groundwork to unify results across different paradigms for a complete account of selective attention processing in the brain [10].

Despite the growing body of work supporting evidence that responses across auditory cortex are modulated by attention, translating such knowledge into computational models has been slow. One avenue in modelling has been to explicitly characterize the adaptation mechanism of STRFs. Mesgarani et al. [66] hypothesized that the spotlight of attention works to enhance the separation between task-relevant target stimuli and the distractor background. Thus, the optimal STRF can be modelled as the filter that gives the highest discrimination between the neural responses to target and background acoustics, resulting in a deterministic linear system that can apply gain to physical features of the auditory input. In this framework, selective

attention can work in a multitude of ways by enforcing different constraints based on perceptual goals, e.g. when listening for short dripping sounds to find the source of a water leak, optimization cost would be increased for slow temporal dynamics, or when attending to a male in a busy room full of children, lower pitches would be enhanced. While relatively simple, this model provides a powerful account of attentional effects at the single-neuron level. Still, it is limited in its ability to extend beyond orienting attention to physical properties of sound (e.g. attend to a class of sounds as opposed to a specific exemplar) and is invariant to task structure due to its implementation. In the last example, if the task were to ignore the male, the adaptation result would not be guaranteed to be different from that for the attend task, as the model separates two signals (male, children) without a conceptual knowledge of task demands (target/distractors).

Recognizing these limitations, Carlin & Elhilali [67] proposed a framework to account for an explicit notion of foreground and background, assigning binary labels to distinguish target sound segments from reference segments as defined by a behavioral task. The addition of task structure to the model resulted in opposite adaptation patterns when the task was switched between reward (foreground) and evasion (background), in line with observed neurophysiological responses at the level of primary auditory cortex in behaving animals [56]. The model was expanded to allow for object-based attention selection, which can "focus attention" on simple abstractions based on physical properties of sound, rather than the acoustics themselves. For instance,

attending to speech as a sound class (regardless of specific utterances and who the speaker is) requires ignoring details of specific acoustic instantiations and responding to abstracted representations of speech that distinguish its characteristics from those of other classes (or objects). The authors modelled such object-based selection as constraints on magnitude and phase profiles of the spectro-temporal dynamics of sound, and provided simulation results to show that modelled STRFs sharpen and orient to target modulations in line with reported physiological effects [55]. Future research is necessary to unify the feature and object-based attention models, and provide neural recording data that better account for attention to complex abstractions of sound.

Another body of work modeled selective attention in a more abstract way by incorporating the attentional gains observed in neurons across physiological experiments into computational models implementing various components of auditory scene analysis. Kalinli & Narayanan [68] extracted the gist of an auditory scene from the biologically motivated acoustic features used in their model of salience [32], and employed a neural network to automatically learn optimal gains given specific tasks, such as scene classification. Patil & Elhilali [69] implemented the hypothesis that attention acts as a prior in a Bayesian representation of the information from the senses [7]. This model used a two-stage computational framework for recognition of acoustic scenes: a feature-extraction stage that mimicked processing in the auditory pathway from cochlea to primary auditory cortex, and an object-mapping stage that performed classification of features into scene types. Top-down attention worked at

both the feature and the object level by applying gains onto the spectro-temporal filters that extract the features, and by adjusting the parameters of a scene classifier to optimize detection of the target scene.

The studies described thus far have taken a forward approach to modelling attention: given the sound input, they predict neural responses and compare the model output with brain responses. Some recent studies have taken a reverse approach to characterizing attention by reconstructing the sound input from recorded neural signals and comparing the reconstructed acoustic waveform with the input to illuminate the aspects of the soundscape that are most prominently represented in the recorded cortical area. While employing regression methods to reconstruct the sensory input from neural recordings is not new, the potential of this paradigm to study the effects of attention has only recently been used to demonstrate exciting results. Mesgarani & Chang [9] reconstructed the spectrogram of the input from intracranial recordings to show that neural representations code salient acoustic features of sound; the reconstruction correlated most strongly with spectro-temporal areas of high energy in the attended source. Further, Ding & Simon [65] reconstructed the input sound envelope from MEG recordings to show that it correlated more closely with the attended speech than the unattended speech in a scene of competing concurrent speakers. This set-up has been extended to reconstruct the attended speech from noisy single-trial EEG recordings [70, 71], an especially important development for the EEG domain where noise reduction techniques coupled with averaging a high number of trials are

typically necessary to estimate the neural signal. With this established framework, biologically plausible models are being designed to reconstruct the input sound from neural recordings, using dynamic state-space models [72] and deep neural networks [73], extending our understanding of attentional gain at the systems level.

2.3 Validation of auditory attention models

While eye-tracking data provides objective evaluation metrics for vision models, attention models in audition have suffered from a lack of clear salience metrics. Most attention models mentioned in this review use their own validation data and metrics, ranging in scope from single-neuron activity to human responses or carefully selected sound events or scenes with attended or salient "ground-truth" determined conceptually by the experimenter. Unfortunately, there is so far little consensus on the best way to probe effects of attention on auditory perception, whether it is task-directed or purely based on salience.

With the first auditory salience models, behavioral experimentation was employed merely to illustrate that the model could detect objectively salient events, such as an animal call amidst pure noise. These studies had subjects choose the more salient of two presented scenes [33, 36, 74], and used a variety of natural environmental sounds as the salient events. Later models adapted more sophisticated paradigms where

the background had a predictable structure, and the task was the detection of the salient event, which had a deviation from the predictable pattern [38, 46], such as a violin note popping out of a stream of piano notes. While these efforts provide a structured way of investigating the precise characteristics of salience perception, their artificial structure limits their account of salience in realistic settings. Attempts at using unstructured natural soundscapes to probe the perception of auditory salience are being made, where subjects listen to real recordings and denote by an interface the time instances they think are salient or interesting [75]. However, unlike the visual domain, in which automatic eye saccades can be rapidly recorded for many scenes, the auditory method is not only much slower and inefficient, but suffers from conscious decision-making, and arguably does not represent purely bottom-up processing as well as its visual counterpart. An intuitive and objective ground-truth dataset for auditory salience would probably lead to a significant increase in modeling efforts, both in designing specialized computational systems that perform robust and efficient computations that can be incorporated into real-time naturalistic applications, and in comparing the performance of various mechanisms hypothesized to underlie neural attentional processing.

On the neural front, single-unit recordings from cats, monkeys and ferrets provide the most direct access to effects of attention on neural activity in the auditory system. While very informative about neural correlates of attentional modulation on brain networks, they are costly to perform, too invasive for human research and are

limited in the amount of information that can be extracted about the intricate cortical networks engaged in auditory perception. They are also restricted to relatively simple or constrained behavioral paradigms that can be used to train animals in a laboratory setting.

The closest correlate to single-unit recordings for humans is electrocorticography (ECoG). Though highly invasive and applicable only to neurosurgery patients, this technique uses electrode grids placed on the exposed brain to investigate attentional modulation of cerebral cortex using rich and complex stimuli. By contrast, MEG and EEG offer non-invasive alternatives that are applicable to a more general population, even though they lack the spatial resolution of EcoG, and are more susceptible to artefacts. Unlike other behavioral measures, MEG and EEG also allow direct insights into neural processes without engaging explicit perceptual decision. However, analytic techniques need to be improved to balance the elimination of noise and preservation of neural information about attentional and perceptual states of subjects, especially in complex sound environments [76]. Further, particularly in studies of bottom-up attention, a common experimental design is such that the subject is instructed to ignore the auditory input and remain engaged in a visual task such as watching a silent film or reading a book. This paradigm is vulnerable to top-down attentional confounds in the absence of distracting auditory stimuli or sufficiently engaging visual tasks.

2.4 Applications of auditory attention models

Aside from providing important contributions to theoretical neuroscience, models of attention play a significant role in a large variety of engineering applications. Particularly, performance on tasks for which humans effortlessly outperform computers could be improved with attention mechanisms, where the attention component would act as a filter to guide computational resources to areas of maximum information, effectively reducing system noise by ignoring irrelevant parts of the scene. One such task is speech and sound recognition: although a trivial task to perform for humans, automatic recognition suffers from significant performance degradation in noisy environments. Some of the surveyed modelling studies have demonstrated various ways in which attentional mechanisms could work to improve existing recognition technologies. Feature-based approaches make use of the feature-extraction schemes of salience models as a way to get a perceptually informative representation of sound input. This representation can be used to detect prominent syllables from speech [32] or as an intermediate step for traditional speech feature extraction and recognition, or fed directly into a clustering mechanism for sound or emotion classification [69, 77]. Top-down taskbased adaptations have been incorporated in attention systems by modelling the attentional gain as weights in the classifier to optimize performance based on specific task goals [77, 78], or as a separate cognitive model

deciding which speaker to attend to among competing sources [78]. A more holistic attention mechanism has instead used the goal-directed adaptation framework of physiological STRFs as a pre-processing stage to speech recognition, by enabling the separation of the target speech stream from the distractor soundscape it is embedded in [79]. The attentional filter provides significant gain to the target speech while being robust to previously unseen noise types. This system was further generalized to use model STRFs optimized for the task, where STRFs are designed as two-dimensional filters, with their parameters estimated from training data [80]. Parametrizing the STRFs allowed for greater flexibility in implementing plasticity. While the authors demonstrated that this model resulted in better identification of speech in noise, the underlying framework can also be applied to a variety of auditory scene analysis problems by training STRFs for specific tasks.

The beneficial effect of attention has also been incorporated into numerous computational auditory models; we give some illustrative examples here. One computational system incorporated both bottom-up and top-down components to mimic human attentional orienting in a busy acoustic environment, allowing a soundscape designer to evaluate how the sound in planned urban environments might affect people [81]. A bottom-up attention mechanism specifically designed for efficient auditory surveillance demonstrated powerful detection of alarming sound events such as gunshots and screams in natural scenes [82]. It has been suggested [47] that attention models are of great importance for improvements in sonification systems aimed at converting

information into sound (e.g. as aids for the blind). An integration of bottom-up and top-down modelling techniques replicating processes in the auditory pathway was demonstrated to improve sound localization in reverberant environments [83]. Auditory salience has been demonstrated to be an effective criterion for compression to reduce data size while retaining meaningful segments of large datasets of sound [84] and video [85]. Salience extraction has also been used as an abnormal sound detection mechanism for temporal signals, and generalized to lung sounds to use for finding medical abnormalities [86].

Finally, auditory attention models are an important component of audiovisual models and applications. In recent years, the necessity of incorporating auditory salience information in visual attention models is becoming increasingly recognized. This has led to the emergence of models using auditory salience direction to guide visual attention [87, 88], along with audiovisual models where the two domains have equal weight in determining attentional orientation [89]. These models show better performance than visual-only salience models in predicting eye gazes in videos.

Mechanisms of multimodal attention are especially crucial in efficient designs of robotic systems [90] and braincomputer interfaces (BCIs). EEG being the most portable method by which brain signals can be recorded, models extracting cognitive information from EEG recordings are of particular significance for BCI systems. The surveyed stimulus reconstruction mechanisms that demonstrate the ability to detect who the subject is listening to have significant implications for powerful nat-

uralistic BCIs. It is of particular interest that these methods are being optimized to use fewer electrodes and faster paradigms to achieve more portable, real-time interfaces [71]. Artificial intelligence systems need attentional filters to select sensory input to process in a goal-oriented manner, and to be able to adapt to unpredictable natural environments. Attention mechanisms have been modelled in various robot and machine-sensing applications [91, 92]. However, these systems use platform-specific definitions of salience and attention, and do not have a direct correlate in the purely computational attention models described here. The computational modeling field has seen significant advances since these robotic sensory designs. Exploring the applicability of new models in robot perception can provide valuable direction to future models, and as computational architectures develop refined biologically plausible mechanisms, human-like robots will become a closer reality.

Chapter 3

Investigating bottom-up auditory attention

3.1 Introduction

Sounds in everyday life seldom appear in isolation. We are constantly flooded with a cacophony of sounds that impinge on our ears at every instant. Our auditory system is tasked with sorting through this sensory flow, to attend to and identify sound objects of interest; all while ignoring irrelevant distracters and ambient backgrounds - a phenomenon referred to as the "cocktail party effect" [93]. A key process in parsing acoustic scenes is the role of attention, which mediates perception and behavior by focusing both sensory and cognitive resources on pertinent information in the stimulus space. At a cocktail party, we can tune out surrounding sounds to listen to one

specific conversation, but the shattering sound of a waiter dropping a tray of glasses will nonetheless cause us to pause to attend to the unexpected event.

Attention is not a monolithic process [5]. It can be modulated by "bottom-up" sensory-driven factors, "top-down" task-specific goals, expectations, and learned schemas; as well as "lateral-based" behavioral history and reward [6]. It refers to a process or group of processes that act as selection mechanisms and allow the sensory and perceptual systems to form a processing bottleneck or focus cognitive resources on a subset of incoming stimuli deemed interesting. In the case of purely "bottom-up" attention, the selection process is driven by sensory cues that orient our attention to interesting events in the environment. It is guided by inherent properties of an event that cause it to stand out with respect to surrounding sounds, regardless of the listener's goal or task at hand.

Some stimuli are inherently conspicuous and pop out amidst certain backgrounds. The study of bottom-up attentional effects is ultimately an investigation of physical attributes of sensory space and integrative mechanisms that allow regions of this space to become salient. In vision, bottom-up attention has been likened to a contrast match concept [94]. Visual elements that differ along modalities of color, intensity, orientation, size and depth (among others) are shown to affect visual search [23], and bias eye fixations in natural scenes [95]. The synergy between the physical structure of a visual scene and saliency-based selective visual attention is a complex one [96]; but has nonetheless been translated into successful mathematical implementations

[97] based on contrast analysis of spatial scales [74], local geometry [98] or spectral contrast [99, 100] using a variety of measures including information entropy [101] and natural statistics [37]. Similar approaches have been explored in the auditory modality with limited success. Adaptations of the visual saliency map have been introduced by considering the time-frequency spectrogram of an audio signal as an "auditory image" upon which saliency mechanisms can operate [44]. This architecture has also been extended to extract attributes better suited for the auditory domain such as a pitch [32, 33]. However, these models remain constrained by the limitations imposed by the visual domain in computing within-feature and across-feature competition for attention; limitations that do not exist in the auditory domain [102]. The nature of sound as a time-evolving entity cannot be captured by spatial processing. There have been attempts to remedy this problem by changes to the procedure of computing saliency after feature extraction, but the methodologies used are still adaptations from vision mechanisms [35, 36]. In this work, we discard the traditional framework of computing a spatial saliency map, and employ psychoacoustical experimentation and computational modeling to build a saliency extraction mechanism that broadly mimics processes that are hyphothesized to take place in the auditory pathway.

Although no evidence has been found for a dedicated auditory saliency map in the brain, the well researched mechanisms of deviance detection in the auditory pathway could be potentially implicated in the perception of saliency in audition. The neural correlates of these mechanisms have long been investigated, leading to the birth of

multiple theories [103, 104]. The recent theory of "predictive coding" [39] provides a unifying framework to encompass some of the previously competing theories under the umbrella of an overall Bayesian brain hypothesis [105, 106]. The Bayesian brain uses generative models to predict sensory input, adjusting its internal probabilistic representations based on novel sensory information. In this setup, predictive coding corresponds to minimizing error between bottom-up sensations and top-down predictions, with the corresponding mismatch signaling the detection of a deviant. There has been considerable support for the theory of prediction-based deviance detection in the auditory domain as the best explanation of neurophysiological observations from electroencephalography (EEG) studies employing simple repeating tones and sound patterns [39, 107]. However, there has been no proposal of an explicit tie between this framework and bottom-up attention in complex natural soundscapes. In this work, we aim to bridge this gap by asking whether the predictive-coding theory can provide an explanation for auditory saliency. To this end, we define a salient auditory event as one that deviates from the feature regularities in the sounds preceding it. In the cocktail party example, the salient shattering glasses would differ from the ambient sounds in acoustic attributes such as timbre, intensity, and location.

We conduct human behavioral experiments to gain psychophysiological insight into the dimensions of auditory saliency and their interactions. In the visual domain, the primary method of obtaining a human ground-truth for the saliency measure is to record eye movements while free-viewing images [48, 108]. However, tracking the

orientation of the attentional spotlight in audition is challenging. Kayser et al. [44] have used a paradigm where they ask subjects to compare which of the two presented sound clips sounds more salient. Kim et al. [45] let subjects listen to recordings of a conference room setting and indicate locations where they "hear any sound which you unintentionally pay attention to or which attracts your attention", further defining salient locations as the ones that were indicated by nearly all subjects. Both studies compare the human experiment results with their computational models, but neither tackles the problem of quantifying the effect of specific auditory features or their interaction on saliency. Here, we follow a similar experimental approach by probing stimulus-related attentional perception using single sound clips, and asking listeners whether they heard a salient event. This paradigm allows us to construct structured full-factorial experiments that can map interactions between features with high statistical power. Although this paradigm is not free from top-down effects on attention, it has been argued that it can successfully account for bottom-up attention effects [109].

The current work is guided by the hypothesis that as sounds evolve in a multi-dimensional feature space, regularities among features are tracked, and deviations from these regularities are "flagged" as salient. A broad range of natural stimuli is used to shed light on the conspicuity of and interactions between the dimensions of pitch, timbre, intensity, and timing in busy acoustic scenes. These perceptual features encapsulate much of the information that is extracted from the cochlea to

mid-brain [110]. A limited number of studies have established the existence of two-way interactions in the perception of some of these features [111, 112]; however, the extent of these interactions pertaining to attention is yet unknown. Here, we probe the effect of these features on auditory attention in a series of full-factorial pyschoacoustical experiments, in an attempt to map the entire interaction space. The same paradigm is used in each experiment, with different modalities of stimuli (musical tones, bird sounds, speech). Short sound clips containing temporally overlapping tokens of sound (e.g., musical note, word) varying in a small range of feature parameters form the scene's "background". Only one token in the scene, the "foreground", is manipulated according to factorial conditions to have a larger feature difference than the background tokens, and could appear at any moment in the scene. Upon presentation of a scene, the subject reports whether they heard a salient event. Results of the behavioral experiments demonstrate the principles governing the influence of acoustic properties on stimulus-induced attention.

In line with our stated hypothesis, we develop a computational model providing an implementation of predictive-coding to test for the first time whether the Bayesian brain framework can explain the perception of auditory saliency revealed by our behavioral experiments. The model analyzes the evolution of sound attributes over time, makes predictions about future values of sound features based on past regularities, and nonlinearly integrates any flagged deviances to yield a unified estimate of saliency over time. The output of this computational model is contrasted with the

pyschoacoustical findings from the behavioral experiments, providing a springboard for exploring the role of inference, predictive representations, and nonlinear sensory interactions in mediating attention in audition.

3.2 Methods

3.2.1 Experiments

Healthy subjects with normal hearing participated in the experiments with informed consent, as approved by the institutional review board at the Johns Hopkins University, and were compensated for participation. Subjects were Johns Hopkins University students and scholars with an average age of 22.6 (number of subjects were Exp. I: 13, Exp. II: 10, Exp. III: 10). All experiments have the same setup: Subjects listen to short sound clips through Sennheiser HD595 headphones in a sound proof booth and answer saliency-related questions on a computer. All subjects in a given experiment listen to the same trials in randomized order. Each trial is presented only once. Trials consist of a dynamic background constructed by many sound tokens that overlap in time with varying density depending on the experiment (Fig. 3.1. Background tokens are randomly selected from a pool of suitable tokens, leading to unique overall backgrounds in each trial. Backgrounds are manipulated so that there is a uniform distribution of frequencies over time, to minimize coincidental increases in pitch difference between the background and foreground tokens. Control

trials consist of just the background scene, while test trials have one "foreground" salient token in addition to the background. The foreground token differs from background tokens in one or more of the experiment factors (i.e., acoustic attributes of the foreground token). Following each trial, subjects are asked "Does the clip contain a salient event?" and report Yes/No answers without feedback. Each experiment is preceded with a brief training session comprised of 7-12 trials that are similar to experimental trials but with feedback provided about which sound feature is changed in the foreground token. Subjects can adjust sound intensity to their individual comfort level in all experiments, at any time during the experiment.

Subject performance is measured with the d' metric, which accounts for false detection rate along with the correct detection rate. In the calculation of d', the detection rate changes according to factorial conditions (averaged between the repetitions of the factorial condition), however the false detection rate is constant for each subject (average of all control trials for the duration of the experiment, since there is no way to attribute a false detection to a particular factor). For both correct and false detection rates, values of 0 and 1 are adjusted to 0.01 and 0.99 respectively. This adjustment is in line with corrections commonly used for d' measures to avoid infinite values. It is worth noting that similar results are obtained irrespective of the small adjustments to the correct and false detection rates. In the analysis of each experiment, the d' was calculated for each factorial condition for every subject. All performed ANOVAs are fully within subjects, where every feature is treated as a fixed

effect, and individual error terms are used in the calculation of the F statistic. The Benjamini-Hochberg procedure [113] is used to iteratively validate the significance levels for multiple comparisons shown in Tables 3.1 and 3.2.

Although the backgrounds in the trials are not identical, there is a possibility that subjects learn the backgrounds over time because of the limited set of background tokens. It is difficult to obtain speech and bird song data from the same source that have near identical pitches but are unique vocalizations. In the case of music, the number of musical notes is predetermined for each instrument, leading to a limited set of notes constrained in a small range of pitch. However, we examine the difference between number of errors in the first half vs. second half of each experiment, and find no significant difference (Exp. I: $F = 1.44, p = 0.24$; Exp. II: $F = 0.49, p = 0.49$; Exp. III: $F = 0.23, p = 0.64$). Furthermore, results from Exp. III confirm that detection of tokens in the beginning of each trial is low throughout the experiment (Fig. 3.3b), refuting the possibility of meta-learning.

3.2.1.1 Experiment I: Music

The first experiment uses a background of non-melodic natural instrument sounds. Non-sustained single notes from the RWC Musical Instrument Sound Database [114] are extracted for Pianoforte (Normal, Mezzo), Acoustic Guitar (Al Aire, Mezzo), Clavichord (Normal, Forte) at 44.1 kHz. Background notes range between 196-247 Hz (G3-B3). Each token is 1.2 s in duration and amplitude normalized relative to its

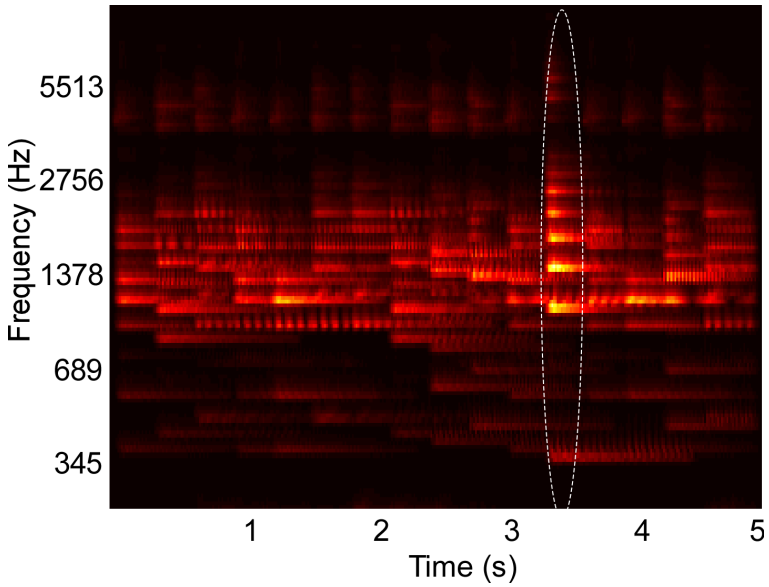

Figure 3.1: Example spectrogram of stimulus used in behavioral experiments. The spectrogram shows overlapping musical note tokens that compose a scene's background. Their pitch and and intensity values are sampled from a constrained distribution of values, emulating a busy scene with natural sounds (Background pitch between 196 and 247 Hz). Listeners cannot perceive any individual note but are able to tell the class of sounds playing in the background. One "foreground" note that varies in pitch (Foreground pitch at 350 Hz) and intensity (6 dB higher than background notes) is introduced at a random location in the scene. In Experiments I and II, foreground tokens only appear in the second half of the scene, while in Experiment III, they can occur at any time. In all experiments, foreground tokens differ from the background in one or more of the following features: Pitch, intensity, and timbre. In the example shown in the figure, timbre was not varied. All tokens were clavichord notes.

maximum with 0.1 s onset and offset sinusoidal ramps. 4 sequences of consecutive tokens, randomly chosen for each trial, are combined with 0.3 s phase delay to form a 5 s dynamic background. Each test trial has one foreground note at 2 or 6 semitones (278Hz-C#4, 350Hz-F4) and 2 or 6 dB higher than background, added at a randomly chosen onset time between 55-75% of the trial length. The resulting experiment design is (Pitch * Intensity * Timbre-foreground * Timbre-background) 2 * 2 * 3 * 3. Each

test condition is repeated 8 times (with non-identical backgrounds). 25% of trials are control trials. Control trial tokens vary in the same range of pitch and intensity as background tokens of test trials. One third of control trials use Pianoforte, one third Acoustic Guitar, and one third Clavichord.

The instruments in this experiment were manually selected such that the they are sufficiently distinguishable from each other, but not so much that listeners with normal hearing and musical training would detect each different note, as determined by short pilot investigations with few listeners. The difference levels for pitch and intensity were similarly set manually to result in a difference that can be definitely heard if one listens for it, but might be missed if not paying attention. The factor levels for subsequent experiments were also set with these criteria.

Experiment I-2 An additional experiment is performed to validate the main effects of musical instruments on the perception of saliency. In this experiment, pitch (5 and 10 semitones higher and lower than the background mean), intensity (7 and 10 dB higher than the background tokens), and timbre are tested separately. Sustained single notes from the RWC Musical Instrument Sound Database [114] are extracted for Harmonica, Violin, Flute (Normal, Mezzo for each) at 44.1 kHz, and downsampled to 16 kHz. Background notes range between 587-740 Hz (D5-F#5). Each token is 1 s in duration and amplitude normalized relative to its top 10%th value with 0.5 s onset and 0.01 s offset sinusoidal ramps. Tokens overlap every 0.5 s, forming two sequences. The foreground token varies in only one of the dimensions with respect to

the background, and is placed at a random onset between 50-80% of the trial length. In each trial, subjects are presented two sound clips, one or none of which contains a salient token. The subject is asked "Which clip contains a more salient event?" and is presented the options "Clip 1"/"Clip 2"/"Equal". Each condition is repeated 4 times, with additional 20% control trials.

3.2.1.2 Experiment II: Nature

The scene setup of this experiment is a busy natural forest environment with singing birds. Natural song recordings of two different Common Yellowthroats, and one MacGillivray Warbler are obtained from the Macaulay Library (http://macaulaylibrary.org, reference numbers: 118601, 136169, 42249). Individual calls at approximately 4.9 kHz pitch and 1.3-1.5 s length are manually extracted at 44.1 kHz. Recordings of wind and water sounds are added to every trial to reduce signal-to-noise ratio, and make the task more challenging while retaining the "natural" scene set-up. Due to unavailability of higher pitched calls from the same bird, background tokens are manually shifted 3 semitones higher with Adobe Audition to be used as foreground tokens. Additional foreground songs with 0 semitone pitch difference are also used, with a change in another attribute (intensity or timbre) following the factorial experimental design. Tokens are amplitude normalized relative to their top 5%th value. Recordings of water and wind sounds (one track for each) are each normalized to have the same peak amplitude as the combined background, and further added to the background.

The foreground token is 2 dB or 8 dB higher than the background. Three sequences of bird calls with 0.5 s phase shift are added for a total duration of 6 s. The foreground token onset is randomly chosen between 58-68% of the trial length. Each individual background token is used at most two times within the same trial. The resulting experiment design is (Pitch * Intensity * Timbre-foreground * Timbre-background) 2 * 2 * 3 * 3. Each condition is repeated 8 times with additional 25% control trials. Control trial tokens vary in the same range of pitch and intensity as background tokens of test trials. Each third of the control trials uses one of the three bird sounds in this experiment.

3.2.1.3 Experiment III: Speech

The background in the third experiment emulates a party scene where one can perceive that people are speaking, but cannot make out what is being said. A noisy telephone conversation recording of two female Japanese speakers is selected from the CALLHOME Database (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96S The choice of Japanese in this experiment is deliberate to ensure non-linguistic interpretations from our non-Japanese-speaking listeners. Further, unlike in Exp. I, one cannot make out individual tokens even while actively attending to them, due to the high level of word overlap and noise in the source recording. 56 words in the 175-233 Hz (F3-A#3) range and of 0.5-1.2 s length are manually extracted at 8 kHz to be in the background. Each word is allowed to appear at most twice in one trial. Each

token is amplitude normalized with its top value and applied a 0.05 s long onset and offset ramp. The background consists of a combination of four sequences of tokens with no delay. Foreground tokens are 10 dB and 13 dB higher from the cumulative background. A foreground token consists of a sample from a selection of 12 words with approximately 8 semitone difference from the background between 349-369 Hz (F4-F#4), each 0.5 s long. Additional foreground words with 0 semitone pitch difference are also used. The foreground onset is also manipulated by placing it in one of four 1.25 s long quadrants of the 5 s long trial, hence probing the effect of timing of foreground on perception of saliency. The resulting experiment design is (Pitch * Intensity * Timbre-foreground * Timbre-background * Time) 2 * 2 * 2 * 2 * 4. Each condition is repeated 4 times, 7.25% are control trials. Control trial tokens vary in the same range of pitch and intensity as background tokens of test trials. 60% of control trials use one speaker, while 40% use the other speaker.

3.2.2 Computational Model

3.2.2.1 Computation of sound features

The model starts by extracting acoustic attributes of the incoming signal with a focus on intensity, pitch and timbre (Fig. 3.2. Intensity is derived from an estimate of the signal's temporal envelope, extracted from the magnitude Hilbert transform, Butterworth filtered with w_c= 60 Hz, n=6. Pitch and timbre are extracted from

the sound spectrogram, which is computed with 1 ms frames. The spectrogram computation mimics the processing known to occur from the cochlea to the midbrain: Using a bank of 128 constant-Q bandpass log-scale filters, followed by high-pass, compression, and low-pass filtering then spectral sharpening following the model of [34]. Pitch is extracted from a harmonicity analysis of spectrogram spectral slices, following a template matching approach [115, 116]. Only pitch estimates with a good match to the template are retained, and further smoothed using a median filter with a 5-sample window. Timbre is a more abstract, less quantifiable attribute, than pitch or intensity. Earlier work argued a close correspondence between timbre perception and spectro-temporal details of sound events [117]. Here, we follow the same premise and first augment our feature space directly with the channels of the spectrogram. In addition, we extract bandwidth information that highlights broad vs. narrowband spectral components; along with temporal modulations that follow dynamic changes of sounds over time. The temporal response of each spectrogram channel is analyzed using short-term Fourier transform with 200 ms windows with 1 ms overlap. Spectral slices of the spectrogram are processed further using Gabor bandpass filters with characteristic frequencies logarithmically distributed between 2^{-2} and 2^4 cycles/octave to extract bandwidth details [34]. The top 64 and bottom 64 channels of the spectrogram are treated as separate features in subsequent processing as high and low frequency spectrum features. The full mapping consists of a 167-dimensional tensor. Finally, each computed feature is further binned using 200ms

windows, such that the mean of the window is assigned to every sample in the window.

3.2.2.2 Deviance detection on feature streams

Following the framework of predictive-coding, each of the model features (envelope, harmonicity, and each frequency channel in high-frequency spectrogram, low-frequency spectrogram, bandwidth, temporal modulation) is separately tracked over time by a Kalman filter [118], which is a linear dynamical system that estimates the channel's state based on measurements over time, by minimizing the least square error between the predicted and observed input. The Kalman filter is used because it is efficient, versatile, and simple to implement and interpret. At each feature channel, clustering on a short segment at the start of the feature decides the regularities to be predicted for that feature. Each regularity stream is tracked with a separate Kalman filter, leading to multiple predictions for incoming values among each feature. If a feature does not fit any of the Kalman predictions, it produces a spike at that instant, signaling a deviant; and a Kalman filter for this novel value is initialized. Filters that are not updated for one second are reset. The match between the input and prediction is determined by a dynamic threshold that depends on prior prediction accuracy. Consequently, if predictions have been matching the input for some time, the expectation is that predicted values will keep being encountered, leading to a decrease in the fit threshold. As the dynamical system evolves, a series of spikes are generated corresponding to times of salient events. The amplitude of each spike corresponds

to the difference between the real feature measurement at that time and the closest prediction window. Finally, spike trains from multi-channel axes (e.g. different frequency channels in the high-frequency spectrogram) are grouped together. If there are multiple spikes at the same time instant, the maximum one is recorded.

The underlying linear system for the Kalman filters in our model is:

$$A(t) = FA(t-1) + u(t)$$

$$Z(t) = HA(t) + v(t)$$

where A is the time-dependent state (or feature variable) being tracked. Z is the observed input. u and v are small Gaussian noise perturbations, modeled respectively as:

$$u(t) \sim \mathcal{N}\left(0, \Gamma = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}\right) \qquad v(t) \sim \mathcal{N}(0, \Sigma = \sigma_v^2)$$

The variances of the noise parameters are empirically chosen for each feature; set to $\sigma_w = 0.001$, $\sigma_b = 0.01$, and $\sigma_v = 0.06$ for envelope and pitch, $\sigma_w = 0.00025$, $\sigma_b = 0.0025$, and $\sigma_v = 0.0125$ for spectrogram, bandwidth, and temporal modulation. The state vector and the system matrices reflect a random walk, and can be encoded

as:

$$A(t) = \begin{bmatrix} Z(t) \\ Z(t) - Z(t-1) \end{bmatrix} \qquad F = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \qquad H = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

The number of regularity streams (each represented with a separate Kalman) to initialize for each feature is determined by k-means clustering of the first 125 ms of feature values. The numbers of clusters are selected so that the sum of distances within each cluster is smallest. For each of these clusters, a Kalman filter is initialized as shown below. The initial values for the state prediction error are calculated from the last two sample values of the initialization window: If n_i denotes the sample number at 125 ms, then the initial estimate for the state vector, and its corresponding state prediction error covariance then becomes:

$$\hat{A}(t) = \begin{bmatrix} 2Z(n_i) - Z(n_i - 1) \\ Z(n_i) - Z(n_i - 1) \end{bmatrix} \qquad \hat{\Psi}(t) = \begin{bmatrix} 5\sigma_v^2 + 2\sigma_w^2 + \sigma_b^2 & \sigma_w^2 + 3\sigma_v^2 + \sigma_b^2 \\ \sigma_w^2 + 3\sigma_v^2 + \sigma_b^2 & 2\sigma_v^2 + \sigma_w^2 + 2\sigma_b^2 \end{bmatrix}$$

Next, at every time instance, the model iteratively computes its Kalman gain $K(t)$, and updates its posterior estimate of the state $\hat{A}(t)$ and $\hat{\Psi}(t)$; following the

equations:

$$K(t) = (F\hat{\Psi}(t-1)F^T + \Gamma)H^T(H(F\hat{\Psi}(t-1)F^T + \Gamma)H^T + \Sigma)^{-1}$$

$$\hat{A}(t) = F\hat{A}(t-1) + K(t)(Z(t) - HF\hat{A}(t-1))$$

$$\hat{\Psi}(t) = (I - K(t)H)(F\hat{\Psi}(t-1)F^T + \Gamma)$$

The threshold to determine whether an input value fits into the prediction of a Kalman is an adaptation from [119]:

$$|Z(t) - HF\hat{A}(t)| \leq \sqrt{4(\hat{\Psi}_{[1]} + \sigma_v^2)}$$

where $\hat{\Psi}_{[1]}$ is the first element in the matrix $\hat{\Psi}$.

3.2.2.3 Integration of saliency information among features

The result of Kalman filtering is a set of one dimensional spike signals for each feature, shown in Fig. 3.2 as $x_i(t)$, where t is time, and $i \in [1, n]$ ($n = 6$ in our case). These spikes represent some probability of having a salient event at the time instance in which they occurred; the higher the value, the more likely is saliency. Note that spike amplitudes in each signal reflect relative deviance within that feature and are not globally normalized to values in other signals. We normalize contribution of each feature and nonlinearly model integration interactions with constrained logistic

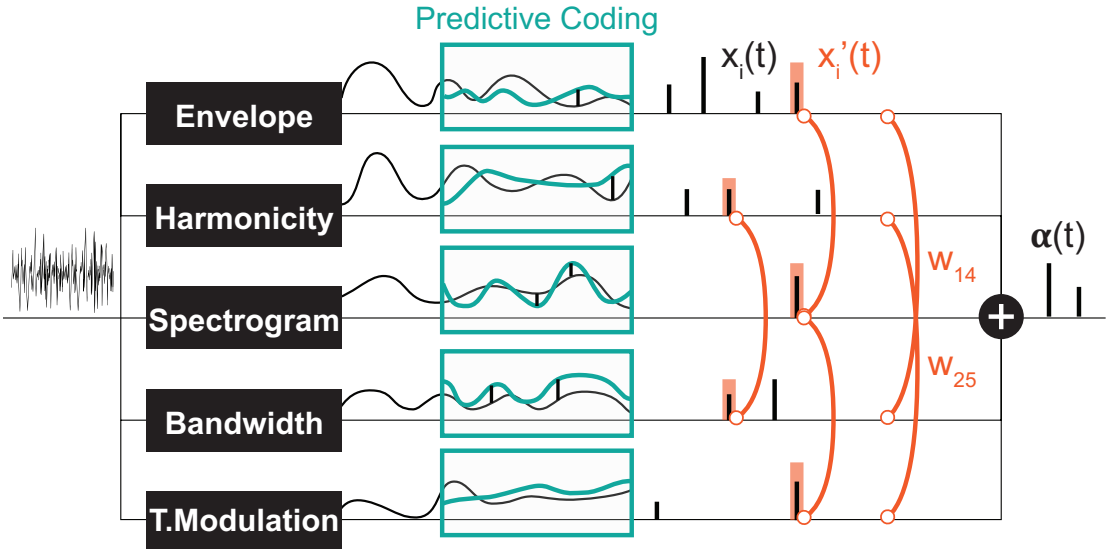

Figure 3.2: Schematic of the computational saliency model. The model is structured along three stages. It starts with an acoustic waveform and extracts relevant features along five dimensions. Regularities within each feature dimension are then tracked used a Kalman-filter to make predictive inferences about deviations from ongoing statistics in that corresponding feature. Detected deviants are boosted according to interaction weights learned using the experimental stimuli, then integrated across feature dimensions to yield an overall saliency estimate of the entire auditory scene. The final values mark salient timings in the scene.

regression, using the stimuli used in our experimental paradigm with their corresponding ground truth about the timings of salient sounds (i.e. timing of foreground tokens).

Let $y(t)$ be a binary variable representing the existence of a salient event in time t. Our objective is to learn a mapping from $x_i(t) \in [0, \infty]$ to $P(y(t) = 1) \in [0, 1]$. An intermediate step in this mapping is boosting the signals (resulting in $x_i'(t)$) with asymmetric interaction weights between feature pairs. This process is illustrated in

Fig. 3.2 and modeled as:

$$x_i'(t) = x_i(t)\Big(w_{ii} + \sum_{\substack{j\in[1,n]\\ j\neq i}} w_{ij} \max_{k\in[-s,s]} x_j(t+k)\Big)$$

w_{ij} are the asymmetric interaction weights between feature i and feature j that we want to find the optimal values of. The window s around a spike accounts for timing shifts due to sampling and is set here to 7 ms. This process is illustrated in Fig. 3.2. The optimal weights w_{ij} are computed using experimental stimuli. The ground truth about deviants in each channel i in these stimuli is:

$$y_i(t) = \begin{cases} 1, & \text{for } t \text{ within salient event duration} \\ 0, & \text{otherwise} \end{cases}$$

We use constrained logistic regression (MATLAB Optimization Toolbox) to map between $x_i'(t)$ and $y_i(t)$. The probability of having a salient event in feature i at time t is determined by:

$$\alpha_i(t) = p(y_i(t) = 1) = \frac{2}{1 + e^{-x_i'(t)}} - 1$$

and the corresponding probability of not having a salient event is:

$$p(y_i(t) = 0) = 1 - p(y_i(t) = 1) = \frac{2e^{-x_i'(t)}}{1 + e^{-x_i'(t)}}$$

With the given binary definition of $y_i(t)$, the probabilities above can be written concisely as:

$$p(y_i(t)|x_i'(t)) = \frac{y_i(t) + (-1)^{y_i(t)} 2^{(1-y_i(t))} e^{-x_i'(t)}}{1 + e^{-x_i'(t)}}$$

leading to the log-likelihood function:

$$\max_{w_{ij}} \sum_t log\Big(\frac{y_i(t) + (-1)^{y_i(t)} 2^{(1-y_i(t))} e^{-x_i'(t)}}{1 + e^{-x_i'(t)}}\Big) \text{ st. } w_{ij} \geq 0$$

Due to the positive constraint on the weights, $x_i'(t)$ is also constrained to be positive, hence limiting the regression to only the positive part of the logistic function. The optimization is performed simultaneously on all features; with clips from all experiments (and their correspondent ground truths) incorporated as training data. For analyses where each experiment is trained separately, each feature is also optimized separately to reduce noise. With the learned weights plugged in, the final output of the entire model is $\alpha(t)$, the likelihood of saliency among time, a value in $[0, 1]$.

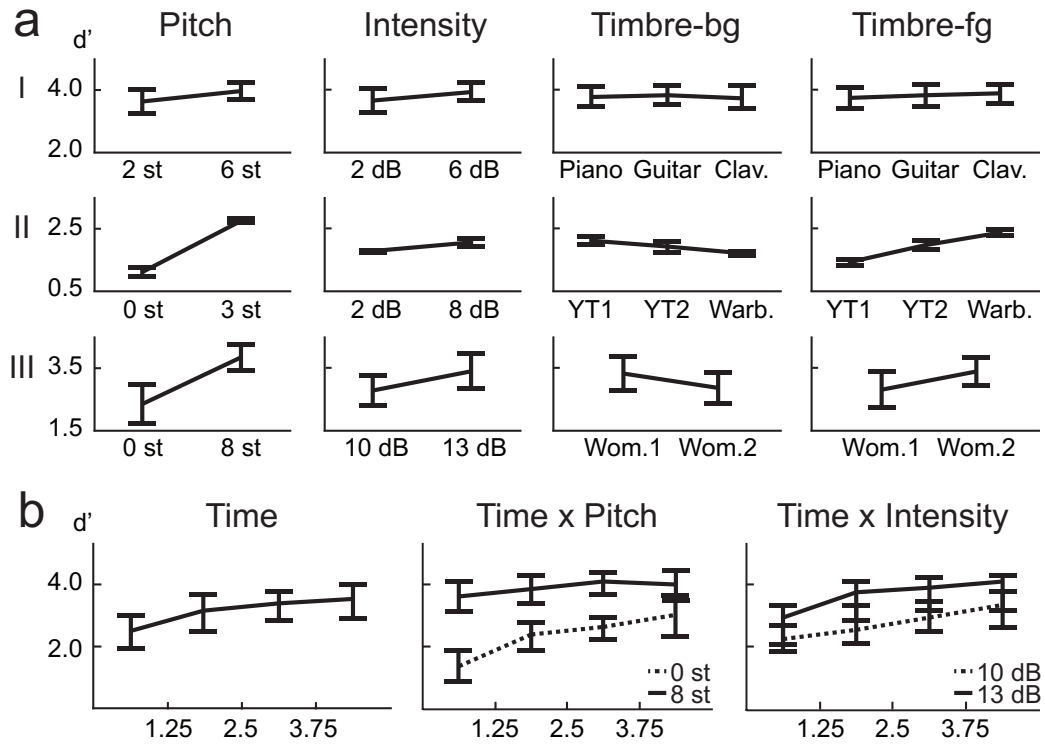

Figure 3.3: Behavioral results. (a) ANOVA main effect trends for all experiments. (b) The effect of the time factor reveals a temporal build-up observed in human detection of saliency. Interaction of time with pitch and intensity are shown. The significance levels corresponding to these plots can be found in Table 3.2.

3.3 Results

3.3.1 Experiments

3.3.1.1 Experiment I: Music

In this first experiment, we investigate the effect of pitch, intensity, and timbre on perception of saliency. Because timbre is a non-numeric attribute, we probe the effect of each musical instrument as a foreground (T_f) and background (T_b) timbre event. Pitch (P) and intensity (I) are found to have significant effects (Table 3.1.

However, neither background nor foreground timbre factors have significant effects. Marginal means (Fig. 3.3a) confirm that the three instruments are indeed relatively close to each other in timbre space; as corroborated by published studies of timbre perception [120]. A follow-up study (Exp. I-2) reveals that the lack of timbre effect is specific to the choice of instruments. An experiment with violin, harmonica and flute (instruments with a wider timbre span [120] shows a statistically significant saliency effect of both foreground and background timbres ($F_P = 4.23$ $p_P = 0.046$, $F_I = 16.44$ $p_I < 10^{-2}$, $F_{T_b} = 8.31$ $p_{T_b} < 10^{-2}$, $F_{T_f} = 4.00$ $p_{T_f} = 0.02$).

3.3.1.2 Experiment II: Nature

Overall, this natural sound experiment is more difficult than the musical notes task (overall d': 1.88 compared to 3.61); but reveals that all four factors have significant effects (Table 3.1. The consistency of effects between Exp. I and II argues against possible ceiling confounds that could have resulted from the musical notes experiment.

3.3.1.3 Experiment III: Speech

In this experiment, we probe the effect of time in addition to the same three attributes tested earlier. Time refers to the placement of the foreground token in the scene, appearing in four possible time-quadrants. All tested factors are found to influence saliency (Fig. 3.3. The trend of the time factor implies that the later a deviant sound is heard in a scene, the more salient it is perceived. There is a significant d'

Table 3.1: ANOVA results of human experiments

Effects	F (p)		
	Music	Nature	Speech
Pitch	**17.76 (<0.01)**	**211.69 (<0.01)**	**103.76 (<0.01)**
Intensity	**14.08 (<0.01)**	**17.57 (<0.01)**	**98.50 (<0.01)**
Timbre-bg	0.63 (0.54)	**8.66 (<0.01)**	**71.21 (<0.01)**
Timbre-fg	2.11 (0.14)	**52.51 (<0.01)**	**29.12 (<0.01)**
P, I	**7.36 (0.02)**	**18.00 (<0.01)**	**134.58 (<0.01)**
P, T_b	0.51 (0.61)	0.09 (0.91)	**19.13 (<0.01)**
P, T_f	1.77 (0.19)	**36.21 (<0.01)**	**12.19 (<0.01)**
I, T_b	1.09 (0.35)	0.98 (0.39)	0.03 (0.86)
I, T_f	0.13 (0.88)	**9.72 (<0.01)**	**11.40 (<0.01)**
T_b,T_f	**13.29 (<0.01)**	**30.21 (<0.01)**	**13.22 (<0.01)**
P, I, T_b	0.28 (0.76)	3.06 (0.07)	**7.03 (0.03)**
P, I, T_f	1.23 (0.31)	0.60 (0.56)	0.39 (0.55)
P, T_b, T_f	**6.77 (<0.01)**	**36.85 (<0.01)**	**33.21 (<0.01)**
I, T_b, T_f	1.57 (0.20)	0.18 (0.95)	**5.60 (0.04)**
P, I, T_b, T_f	0.29 (0.90)	0.24 (0.91)	**7.47 (0.02)**

increase in the first two quadrants of the scene (Bootstrap 95% confidence interval for slope: $(25.6°, 35.8°)$, $p < 10^{-2}$), indicating rapid adaptation to the background (Fig. 3.3b). The trend stabilizes later in time (low difference between last two quadrants; Bootstrap 95% confidence interval for slope: $(-1.1°, 16.7°)$, $p = 0.09$) implying that once standard formation has taken place, detection may no longer be highly dependent on exact timing.

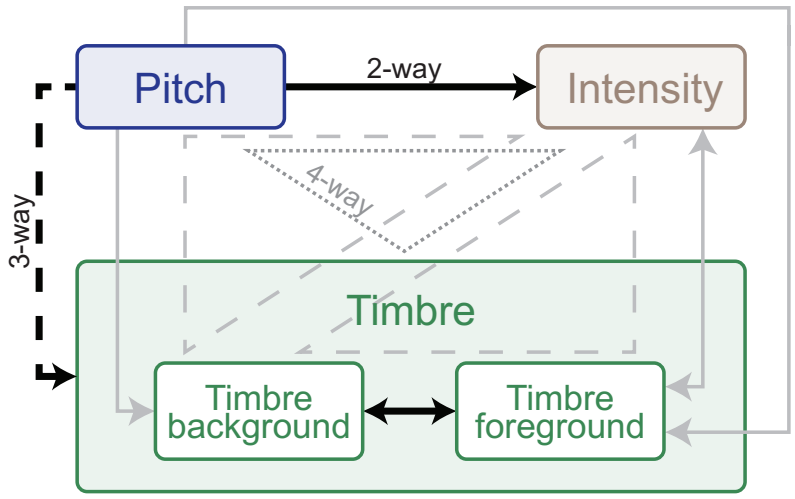

Figure 3.4: Summary of interaction weights based on behavioral tests with human listeners. Solid lines indicate 2-way, dashed lines 3-way and dotted lines 4-way interactions. Effects that emerged in every experiment are shown black, and those that were found in at least one experiment are shown grey. Arrow directions indicate direction of interaction: The origin feature has a relatively larger effect on the destination feature in all experiments. Double-sided arrows indicate that there is no clear weight either way. The weight and directionality of interactions observed are inferred from the coefficients of the fitted model, and are limited by the levels of sound features tested in this study.

3.3.1.4 Interactions

An interaction between multiple factors indicates that the effect of one factor changes according to the levels of the others. Within-subjects ANOVA results, outlining the interactions from all experiments, are shown in Table 3.1. Intensity and pitch have a significant interaction: The effect of intensity is more prominent when pitch difference is low. Although separate timbre components (T_f, T_b) are not significant in every experiment, their interaction is significant; demonstrating that the effect of timbre on saliency stems from the interplay of background and foreground.

Table 3.2: ANOVA results of interactions with the Time factor in Experiment III

	F (p)		F (p)
Time	**42.57 (<0.01)**	Time, I, T_b	2.57 (0.08)
Time, P	**18.90 (<0.01)**	Time, I, T_f	1.76 (0.18)
Time, I	1.12 (0.32)	Time, T_b, T_f	2.77 (0.06)
Time, T_b	2.17 (0.12)	Time, P, I, T_b	2.06 (0.13)
Time, T_f	1.61 (0.21)	Time, P, I, T_f	0.56 (0.64)
Time, P, I	0.87 (0.47)	Time, P, T_b, T_f	0.15 (0.93)
Time, P, T_b	1.43 (0.26)	Time, I, T_b, T_f	0.80 (0.51)
Time, P, T_f	**4.75 (<0.01)**	Time, P, I, T_b, T_f	1.32 (0.29)

Further, while T_f and T_b do not separately interact with pitch in every experiment, the combined interaction PxT_bxT_f does. Thus, one can argue that pitch and timbre have a significant interaction (Fig. 3.4. An interaction between intensity and timbre, and between all four factors, is observed in only one experiment.

Time emerges as an additional significant factor in Exp. III. In one case, the effect of pitch on perceived saliency is found to depend on the length of build-up (Fig. 3.3b). The complete high-level interactions can be found in Table 3.2, corroborating the importance of timing of events for auditory saliency. The higher detection performance when the salient event is later in the scene suggests a notion of accumulation of background statistics over time, in agreement with our hypothesis.

3.3.2 Computational model

The computational model produces a one-dimensional signal indicating the likelihood of salient events over time, corresponding to a "saliency score". The model is run on the same stimuli used in the experiments, with interaction weights obtained by training on the ground truth about salient events. Note that no model training is done to match it to the human ratings. The average model saliency scores for trials with salient tokens are statistically significantly higher than those for control trials (t-test, all experiments: $p < 10^{-2}$). In most trials, the likelihood of saliency is highest during the duration of the actual salient event: I: 61%, II: 78%, III: 92% (Fig. 3.5a). When contrasting the model scores with human ratings, strong correlations are observed (Fig. 3.6a). The saliency scores of repeated factorial cases are averaged for the model. The human responses, mapped to 0 and 1, are averaged over factorial case repetitions, and also averaged between subjects. Statistically significant correlations are found in each experiment, when the model weights are calibrated for stimuli and ground truth from all experiments simultaneously (Spearman's rank correlation: I: $\rho = 0.60$, $p < 10^{-5}$. II: $\rho = 0.63$, $p < 10^{-5}$. III: $\rho = 0.61$, $p < 10^{-5}$.). Higher performance is observed when the model is calibrated for ground truth of each experiment separately (Spearman's rank correlation: I: $\rho = 0.64$, $p < 10^{-5}$. II: $\rho = 0.72$, $p < 10^{-5}$. III: $\rho = 0.80$, $p < 10^{-5}$.). Furthermore, we observe that the model saliency scores increase as the level of saliency increases. The level or strength of saliency of a token is taken as the number of sound attributes in which the foreground is differ-

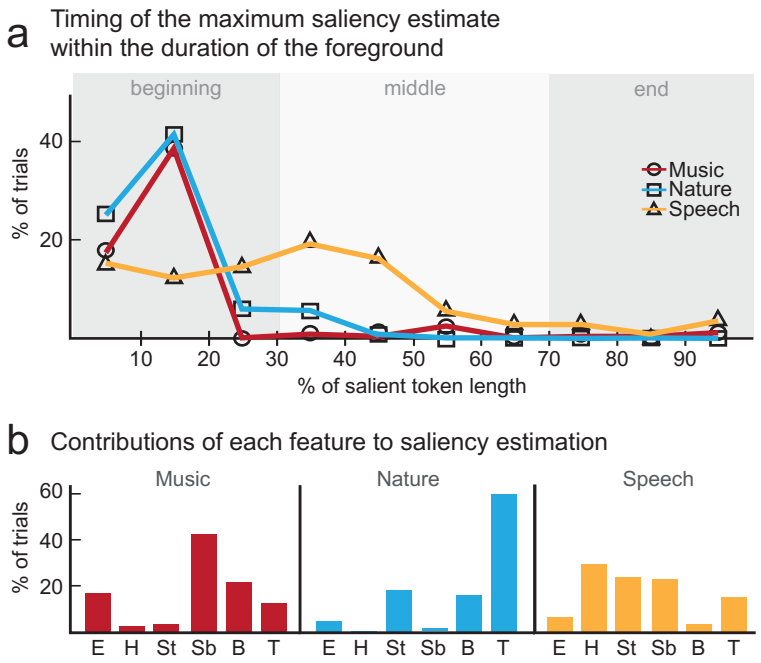

Figure 3.5: Analysis of model results. (a) The time instance where the maximum likelihood of saliency was detected for foreground tokens in the scene. Trials in which the maximum saliency was found outside the duration of the foreground are not included. For musical notes and bird songs, the deviance is detected soon after the token onset. For spoken words, the deviance is detected during the first half of the token onset. In some cases, the model finds the offset deviance instead of onset deviance. (b) Regardless of whether the maximum likelihood of saliency was inside the foreground token duration, the feature that the saliency was detected in is shown. The features are, in order: Envelope, Harmonicity, Spectrogram-top, Spectrogram-bottom, Bandwidth, Temporal modulation.

ent than background. Fig. 3.6c (left) shows the increase in model saliency score as the foreground saliency strength increases (Spearman's rank correlation: I: $\rho = 0.67$, $p < 10^{-5}$, II: $\rho = 0.61$, $p < 10^{-5}$, III: $\rho = 0.64$, $p < 10^{-5}$). The behavior of human listeners is also similar, with average ratings across subjects increasing as strength of saliency increases as shown in the right plot in Fig. 3.6c (Spearman's rank correlation: I: $\rho = 0.83$, $p < 10^{-5}$, II: $\rho = 0.81$, $p < 10^{-5}$, III: $\rho = 0.64$, $p < 10^{-5}$).

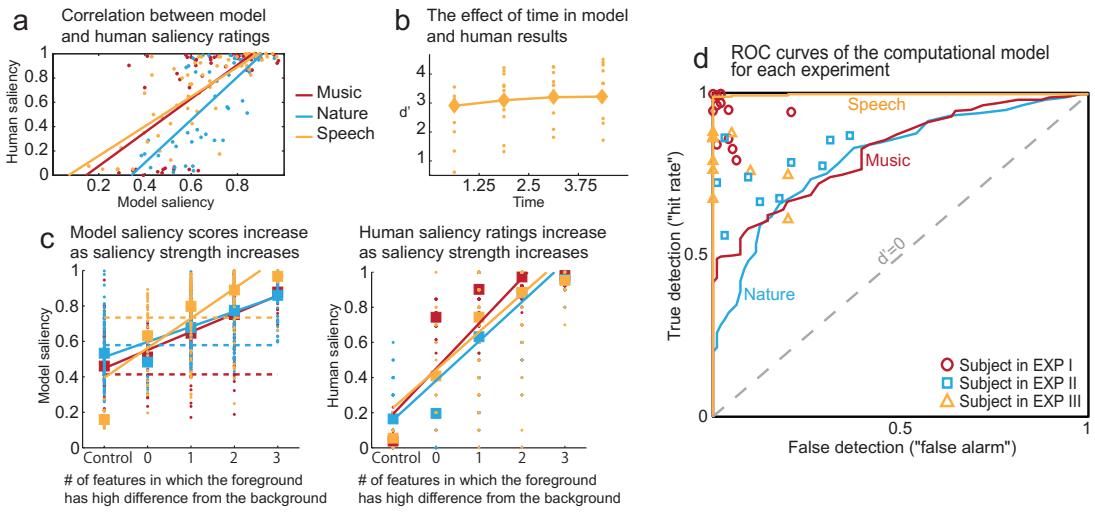

Figure 3.6: Comparisons of human and model results based on saliency ratings and detection performance. (a) Correlation between averaged model saliency scores and human saliency ratings shown for all experiments. Averaging is performed between repeated experimental cases, and also between subjects for the human ratings. (b) The time trend that emerged in the model results for Experiment III. Diamonds show the d' for each quadrant in model results, and dots represent the human responses. We observe a similar trend as in Fig. 3.3b. (c) We show that as saliency increases, the model produces higher saliency scores. This is along the same lines with human results. Control trials have no foreground token; there is no salient event during the scene. Feature level 0 on the x-axis corresponds to a foreground token with low level of saliency. As an example, for Experiment III, this corresponds to no difference in pitch or timbre, but a 10 db difference in intensity. Feature level 1 corresponds to the high level of difference, which is 13 db for intensity in this experiment. Any change in timbre or pitch is also counted as a high difference due to the experimental set-up, outlined in Methods. The dashed lines in the left plot show where the threshold falls for calculating the optimal d'. The separability of control trials from test trials demonstrated here is also reflected in the ROC plot. (d) The probabilistic output of the saliency model leads to a detection curve in ROC space by setting a threshold to distinguish true and false detections. The d' metric can be computed for each point in this space, quantifying performance; d' is 0 when true and false detection rates are equal. We can infer from the curves that the saliency scores of the control trials are most easily separable than the saliency scores of the test trials for Experiment III, and that the performance of the model is closest to humans for Experiment II.

We perform further analysis on the model's behavior and observe that different acoustic features have varying levels of contribution in different experiments; band-

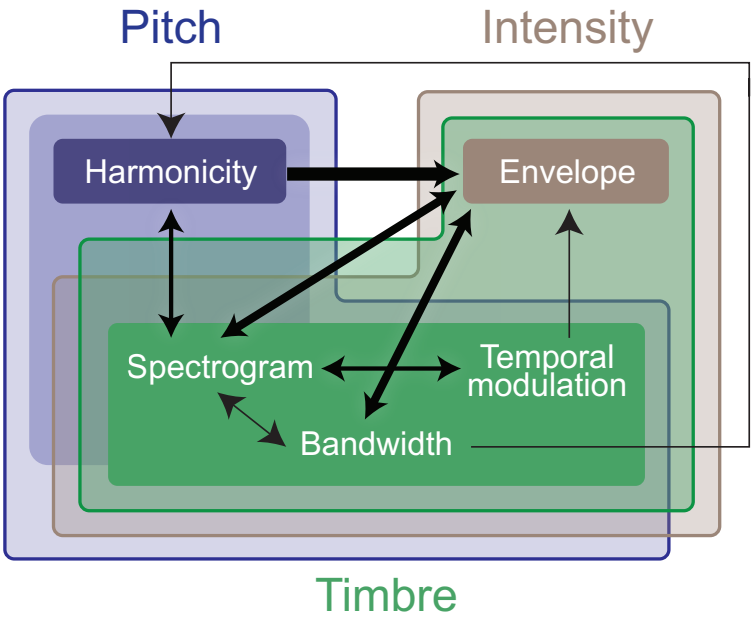

Figure 3.7: Summary of interaction weights that emerge from training the computational model. The model is trained using the same stimuli used in the experimental testing. Thicker lines denote higher weights. An arrow between features indicates that the origin feature of the line boosts the effect of the destination of the line. The different colors indicate the computational features that encode effects of the experimental features, the deeper the color, the stronger the relationship. As in Fig. 3.4, the weight and directionality of interactions in this figure are inferred from the coefficients of the fitted model, and are limited by the levels of sound features tested in the human experiments.

width and temporal modulation appear to be the most effective (Fig. 3.5b). A careful inspection of model feature interactions shows strong similarity with psychoacoustic findings, even though the model interaction weights are trained based on ground truth about deviant events, not on human results. In particular, pitch and intensity have a strong interaction in both human perception and the computational model. The effect of intensity is strongly boosted by pitch; their opposite interaction is weaker. Features capturing timbre have complex interactions between themselves depending on the experiment. It is important to note that the overall interactions observed

reflect the redundancy in the computational features - e.g., intensity is encoded, to some extent, in the spectrogram, and thus bandwidth, therefore these features tend to spike together, leading to likely interactions between them. The observed effects should be interpreted within the context of the feature levels tested in the human experiments.

The probabilistic saliency output of the model can function as a discrete deviance detection mechanism by mapping the saliency scores to a binary classification. The performance of the model as a deviance detector is evaluated with an ROC curve, which maps the discrimination ability of the classifier as true detections ("hit rate") against false detections ("false alarm"). Detection rates are computed for every possible threshold in the range $[0, 1]$ with a step size of 0.001. The resulting ROC curves of the model (with weights from training all experimental stimuli simultaneously) are shown in Fig. 3.6d, along with each subject's performance as mapped onto the ROC space. We select optimal thresholds on the curve based on the d' metric, which quantifies the discrimination ability of the classifier at each location of the ROC space. The average human d' values obtained from our psychoacoustic experiments are: I: 3.61, II: 1.88, III: 2.67. Selecting the thresholds for each experiment that produce the closest hit rate to human results, we obtain d' values of I: 1.11, II: 1.20, III: 3.10. On the other hand, if the model is tuned as an absolute deviance detector (i.e. based on ground truth of deviant events), it yields d' values of: I: 2.29, II: 1.72, III: 4.74. In comparison, the d' values on the same stimuli run through the Kayser *et al.* saliency

model [44] are: I: 0.91, II: 0.78, III: 0.52 (scores correspond to maximum amplitude of the saliency map, parallel to our definition of the saliency score in this study). Moreover, unlike the static nature of previous auditory saliency models, the current computational model reveals a temporal build-up behavior similar to that observed in the speech experiment (Fig. 3.3b). The model d' values corresponding to the four quadrants are: 2.91, 3.10, 3.21, 3.21, illustrated in Fig. 3.6b.

3.4 Discussion

Results from our perceptual experiments reveal an intricate auditory saliency space that is multidimensional and highly interconnected. Some of the observed interactions are not unique to the current study; but have been reported in other contexts of detection, classification and discrimination tasks [111, 112, 121]. The current work paints a more complete picture of the non-symmetric nature of interactions in the context of complex dynamic scenes. Each of the probed auditory attributes (pitch, timbre and intensity) is a complex physical property of sound that likely evokes several neural processing streams and engages multiple physiological nuclei along the auditory pathway. It remains to be seen whether the nature of interactions reported here reflects intrinsic neural mechanisms and topographies of feature maps in the sensory system; or reveals perceptual feature integration processes at play in auditory scene analysis.

The study of bottom-up auditory attention appears to be intimately linked to processes of auditory scene perception and formation of auditory objects. The current work argues for a strong link between tracking statistics of an auditory scene and elicitation of deviance signals that flag salient sounds as aberrant events that would be attention grabbing. This process builds strongly on the notion of predictive inference, and frames the analysis of auditory scenes and selecting events of interest via predictive interpretations of the underlying events in the scene. The saliency processes presented here could be interpreted as signals for marking the reset of the grouping process in auditory streaming; flags of deviant events within an existing perceptual stream; or indicators of initiation of a new auditory object which does not fit within the expected fluctuations of the ongoing stream. Such notion is intimately linked to the concept of regularity tracking as an underlying mechanism for perception in auditory scenes [122], with accumulating evidence that strongly tie predictive models of sensory regularity and stream segregation [123, 124]. Some of the computational primitives presented in the current model could be seen as a shared neural infrastructure that mediates regularity tracking in a sensory-driven way [125], both to provide putative interpretations of the auditory scene as well as flag pertinent events of interest (guided by bottom-up attentional processes). The strong effect of timing on perception of saliency demonstrated by our pyschoacoustical and computational findings further hints to ties between the inference process observed here and the phenomenon of build-up of auditory streaming [126, 127, 128, 129] or its perceptual

stability [130, 131].

The model presented here is a formal implementation of the concept of regularity tracking and deviance detection in the context of dynamic scenes. These concepts have often been linked to studies of auditory attention, though the causal relationship between attention and representations of regularity is still a matter of debate [132]. The physiological bases of deviance detection is commonly probed using mismatch negativity (MMN) [133], a neural marker that emerges as the difference between responses to the "standard" and "deviant" in a stimulus often in an oddball paradigm [39]. The underlying mechanisms eliciting this negativity have been attributed to a potential role of memory [103, 134] or caused by neural habituation to repeated stimulation [104]. A unifying framework for these mechanisms has been proposed in theories of Bayesian inference [39, 135, 136]. The premise is based on the notion that the "Bayesian brain" continuously makes likelihood inferences about its sensory input, conceivably by generating predictions about upcoming stimuli [40]. Predictive coding is arguably the most biologically plausible mechanism for making these inferences, implicating a complex neurocircuitry spanning sensory, parietal, temporal and frontal cortex [137]. The computational framework presented in this study follows the same predictive coding premise to model mechanisms of bottom-up auditory attention. It formalizes key concepts that emerge from our perceptual findings; namely: use of dynamical system modeling to capture the behavior of the acoustic scene and its time-dependent statistics; tracking the state of the system over time to infer evolution

of sound streams in the scene; generating expectations about stimuli that adapt to the fidelity of sensory evidence and lead to a build-up effect of saliency detection accuracy; multidimensional mapping of sensory data that enables integrated cross-channel deviance detection while accounting for complex interactions in this multi-feature space. Kalman filtering is a natural fit for modeling such behavior. It provides an online tool for tracking evolution of states of a dynamical system that reflect past behavior and expected trajectory of the system. In many respects, the Kalman filter is equivalent to iterative Bayesian filtering under certain assumptions [118], and can be implemented using biologically plausible computations in neural circuits [138, 139]. However, the Kalman formulation remains a linearized approximation of the dynamic behavior of acoustic scenes. More suitable frameworks such as particle filtering [140] or recurrent Bayesian modeling [141] as well as non-Bayesian alternatives based on Volterra system analysis [142] need to be investigated to provide a more complete account of the inference process in everyday acoustic scenes.

The use of predictive coding in the model takes a different direction from common modeling efforts of saliency in other modalities, particularly in vision. There is an abundance of models that implement concepts of stimulus-driven visual attention in which the theory of contrast as measure of conspicuity of a location in a visual scene plays a crucial role (see [28] for a recent review). These models vary in their biological plausibility and anatomical fidelity to the circuitry of the visual system, and differ in their focus on sensory-based vs. cognitive-based processes for attentional

bias of visual information. Very few models have explored the role of Bayesian inference in modeling visual saliency. Recent work has started exploring the notions of expectation, predictability and surprise as a conceptual framework for visual saliency [101, 143, 144]. While the notion of "prediction" or predictive coding is implicit in these models, they incorporate many of its conceptual elements and could rely on the canonical circuits of predictive coding that are pervasive throughout processing stages of visual cortex [137, 145]. In parallel, there is greater interest in physiologically probing change detection in vision, particularly its event-related brain potential (ERP) component of visual mismatch negativity (vMMN). vMMN has been described in a number of recent studies over the last decade (see [146] for a review), though it has only been probed using temporal sequences and changing stimuli. Recent findings have also reported somatosensory magnetic mismatch negativity (MMNm) [147] and olfactory mismatch negativity (oMMN) [148], suggesting that MMN is a common framework for change detection across sensory modalities. The ubiquity of deviance detection in sensory cortex raises the question of commonalities among different senses in attentional selection mechanisms; or whether the parallels between audition and other senses are limited to change detection in dynamic sequences and time-dependent signals. Moreover, it remains to be seen whether saliency processes can be fully accounted for by stimulus features that induce pop-out or whether the complex interaction between sensory attributes, global proto-objects, semantic guidance and top-down attentional feedback is necessary to complete our understanding

of bottom-up attention.

Chapter 4

Neural characterization of auditory attention

Attention is a key component of auditory and visual scene analysis [3, 94]. Attention facilitates focus on an object of interest in presence of numerous distractors in a natural environment, allocating cortical resources for detailed processing of the chosen object (top-down, selective attention), all the while maintaining perceptual flexibility to allow important, salient objects to enter the spotlight (bottom-up, automatic attention). Decades of research have contributed to the perceptual and neural characterization of task-directed attention in audition and vision [93, 149], with the underlying assumption that streams or objects that vie for attention have already formed. However, to fully characterize human attention to the natural environment, not only goals, expectations and learned priors [6], but also the neural representations

of the inherent saliency of objects, and the interplay of these internal/external factors must be understood.

The ease of obtaining eye-tracking data for complex natural scenes and the high contribution of early saliency to fixations have been significant factors driving the proliferation of experimental and modeling studies in the visual saliency literature [28]. As there is no direct parallel to automatic saccades in audition, a variety of experimental techniques have been proposed to measure auditory saliency, such as comparing sound clips [36, 44], making salience judgments [38], or actively denoting salient events in free listening [45] or between competing sounds [150]. Ultimately, behavioral measures of bottom-up auditory attention require careful experimental design, with cautious interpretation of results that will likely be confounded by top-down factors [151]. Although attentional orientation cannot be inferred from the ear, electrical or magnetic fields generated by the synchronized firing of large numbers of parallel oriented cortical neurons receiving the same synaptic input can be recorded non-invasively using electroencephalography (EEG) or magnetoencephalography (MEG) respectively. The rhythmic activation of cortical neurons processing current tasks and behavior combined with the self-generated intrinsic rhythms of the brain appear in EEG recordings as a periodic wave [152, 153, 154]. Upon presentation with sensory stimuli, neural circuits processing the incoming information modulate EEG oscillations. These modulations can be teased apart from stimulus-irrelevant oscillations in the temporal domain by comparing different epochs of the stimulus

evoked waveform to a baseline waveform, or in the spectral domain by analyzing the amplitude and phase changes at different frequencies of the waveform during stimulus presentation [155].

Analysis of neural responses to auditory stimuli in the temporal domain has a rich history in the literature. In these studies, stimuli typically contain only a few unique short sound segments that are presented many times, regardless of whether they are part of a task or whether the stimulus is presented to an unattending subject [156]. Further, segments are well separated in time to allow the neural processing to the previous stimulation to finalize. The repetitions of the same stimulus are then averaged to obtain the event-related potential (ERP), with the expectation that intrinsic oscillations are random and their average will approach zero as the number of repetitions increase, while deflections that reflect processing of the stimulus will remain as the ERP "components". ERPs to short sound segments are well characterized, with known components that occur within specific latencies that are thought to be markers of various cognitive processes relevant to sound processing [157]. The component that relates most closely to the study of saliency is the mismatch negativity (MMN), which is observed for sound sequences that involve a high number of repetitions for one "standard" sound interspersed with rare "deviant" sounds [103]. The MMN is believed to be a marker of deviance detection, and has been successfully elicited using a variety of tones and natural stimuli deviations in many auditory features, including frequency, loudness, timbre, duration, and space [39, 133, 158],

and is also observed in other sensory modalities [147, 148, 159]. The advantage of an MMN-eliciting experimental setup is that MMN has been well characterized and can be recorded under completely passive conditions, and thus is a likely candidate measure for purely bottom-up attention (although it can also be modulated by top-down attention, see [132]). However, the cost of the paradigm is that it requires very artificial scenes made of short sounds, many of which are identical, stitched together. Not only does this requirement make it challenging to immediately extend the paradigm to completely natural sound scenes to probe free-listening, it limits the spectral information that can be extracted from EEG oscillations due to its sparse setup in both time and space.

EEG data can also be analyzed in the spectral domain by investigating the modulations of energy and phase across different frequency bands, revealing parallel cortical processes occurring simultaneously. Recent evidence has demonstrated that presented with a rhythmic sensory stimulus, neural ensemble activity fluctuates in a pattern matching that of the attended stimulus, driving the power of oscillations at the stimulus rate [160, 161, 162]. Further, this power is modulated by attention [163, 164, 165, 166]. Entrainment to attended sounds occurs regardless of whether the stimulus is strictly rhythmic [163, 167, 168]; oscillations have been found to phase-lock to the envelope of natural speech [65, 169, 170, 171, 172, 173]. Representations of natural sounds such as animal vocalizations and ambient sounds have also been found to be coded in the phase signature of low-frequency oscillations [174, 175, 176].

The recent studies outlined above have successfully extracted stimulus-specific information from neural recordings to natural, continuous sound environments; however, they have all employed experimental paradigms under the influence of top-down attention. In this work, we asked whether it is possible to record a pure bottom-up attention response with EEG in an experimental paradigm that does not engage the auditory attention of the subject in any way. As the entrainment responses to natural sounds have not been explicitly demonstrated to occur in unattended conditions, MMN is the sole marker of bottom-up attention in the EEG literature. We constructed stimuli using real musical instrument notes, without any melody, but specifically chosen to sound pleasing to the ear, to make it sound more natural despite the artificial structure (same stimulus as the Music experiment in Chapter 3). Our paradigm differs from traditional MMN studies in three ways: (i) the notes, while regularly repeating, occur at a much faster rate than in oddball paradigms (3.33 Hz) and the full note decay lasts 1.2 seconds, (ii) standard notes are non-identical, they vary in a controlled pitch and loudness range to form a dynamic realistic background, (iii) rare salient notes are also non-identical, and deviate from the standards in multiple features (pitch, loudness, and timbre). The latter point is of particular significance, as salient sounds in a real environment are complex and vary among many features, rarely is a salient event identical in every sound property but one.

We demonstrate markers of auditory saliency obtained from an unattending subject paradigm for the first time. While subjects were attending to a visual movie,

cortical oscillations nonetheless entrained to the irrelevant auditory stimulus. This entrainment is found to be significantly more powerful for salient sounds, and is further shown to be modulated by the degree of saliency among a rich acoustic feature space. We confirm that this response represents a bottom-up deviance detection with the presence of the MMN and P3a components of the ERP. We further extract spectro-temporal filters from the neural responses that reveal possible cortical adaptations in processing salient sounds among different acoustic features. Results are contrasted with previous behavioral experiments obtained in Chapter 3.

4.1 Methods

4.1.1 Participants

Twenty-one subjects (9 female) participated in the experiment after giving informed consent and were compensated for their participation. All subjects had normal vision and hearing, and no history of neurological disorders. The experiment was conducted in accordance with the Institutional Review Board of the Johns Hopkins University.

4.1.2 Stimuli and procedure

The auditory stimuli presented to subjects was identical to the Music experiment in Chapter 3, where pyschoacoustic detection results were collected. In the behavioral experiment, subjects listened to 5 second long acoustic scenes and were asked whether they heard a salient sound. Here, subjects were instructed to ignore the sound being played, and their attention was diverted by with a silent movie of their choice for the duration of the experiment.

The stimuli consisted of regularly spaced overlapping musical notes, with a new note starting every 300 ms. We defined the notion of "background" as the set of notes forming the regularity, all of which were the same instrument (timbre) and approximate intensity, but varied within +/- 2 semitones of pitch. Only a single note in every 5 second trial was the "foreground", which was the target in the behavioral experiment [38]. The foreground deviant note differed from the background in any combination of the following acoustical properties: Timbre, pitch, intensity. We selected three instruments (piano, guitar, clavichord), two levels of pitch difference (2 or 6 semitones higher), and two levels of intensity difference (2 or 6 dB higher) to test in a factorial design. The repetition rate and instruments were specifically selected to sound pleasing as if flowing naturally, to resemble natural sounds and avoid a cacophony.

Due to the difficulty of defining timbre on a scale, we characterized timbre difference categorically by testing all 9 combinations of the 3 instruments for standard

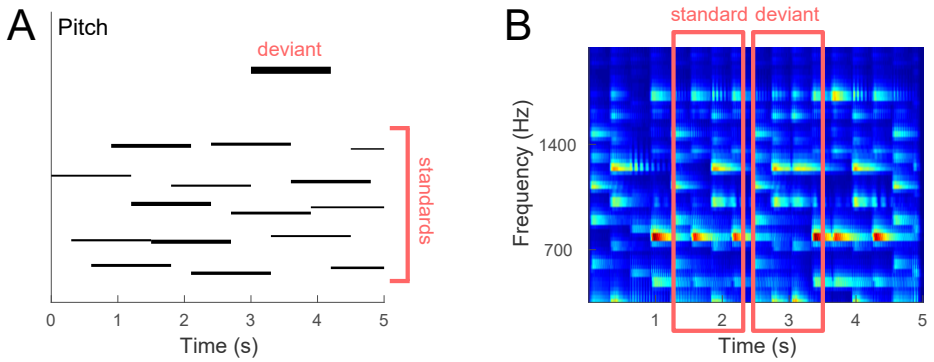

Figure 4.1: (A) Schematic of the stimulus with overlapping musical notes. Line thickness denotes loudness. Standard notes are all of the same instrument and vary randomly in a controlled range of pitch. In each 5 second trial, there is only one deviant note. Deviance in pitch and intensity is shown, all combinations of pitch, intensity and timbre (instrument) are tested. Time of the deviant is random in the latter half of the trial. (B) Example spectrogram of one trial. Standard sections are taken to be the same length as deviant sections immediately preceding the deviant for each trial.

notes (Timbre-background) and deviant notes (Timbre-foreground). This resulted in 3 * 3 * 2 * 2 = 36 trials to test every possible feature deviation. We repeated each feature deviation 8 times, and included control trials where there was no foreground note, for a total of 384 trials (288 deviant trials). Note that each of the 8 repetitions feature a different random dynamic background and different deviant onset time. The order of trials was randomized for each subject.

Instrument notes were extracted from the RWC Musical Instrument Sound Database [114] for Pianoforte (Normal, Mezzo), Acoustic Guitar (Al Aire, Mezzo), Clavichord (Normal, Forte) at 44.1 kHz. Background notes were selected between 196 and 247 Hz (G3-B3). Each note was a total of 1.2 s long and was amplitude normalized relative to its maximum with 0.1 s onset and offset sinusoidal ramps. The deviant note was

placed at a randomly selected time between 2.4 s and 3.8 s from the start of the trial.

The trials were concatenated with 3 s of silence in between. Following every 40 trials, 30 s of silence was inserted to give the subject time to move if necessary. Subjects were instructed to minimize movement and blinking, and keep their eyes at the center of the monitor during the sections were sound was being played. Subjects were seated in a comfortable chair in a dimly lit shielded chamber. Sound was delivered binaurally at a comfortable hearing level via ER-3A plastic tubing connected to ear plugs that were inserted into the ear canal of the subject. Sound delivery was controlled by Presentation software (Neurobehavioral Systems). The movie was presented on a monitor placed approximately 1.5 m away from the subject, except for the case of one subject who had imperfect vision for whom the monitor was placed at a comfortable viewing distance of approximately 0.5 m.

4.1.3 EEG recording and preliminary processing

EEG recording was performed with the Biosemi Active Two system at 2048 Hz. 128 electrodes, plus left and right mastoids were recorded. Four additional electrodes recorded eye and facial artifacts from eye EOG locations, and a final electrode was placed on the nose to serve as reference. However, the nose electrode was found to be heavily contaminated with eye artifacts for most subjects. Therefore, the nose electrode was used only to test mismatch negativity at the mastoids, and the average mastoid reference was used for all further analyses.

The initial processing of neural signals was performed with the FieldTrip [177] software package for MATLAB. Trials were epoched with 1 s of buffer time at both ends, referenced to the left and right mastoid average, and downsampled to 256 Hz. To remove muscle and eye artifacts from the signals, we opted to use independent component analysis as implemented by FieldTrip. ICA components were removed if their amplitude was greater than the mean plus 4 standard deviations for more than 5 trials. Resulting filtered signals were visually confirmed to be free of prominent eye blinks and large amplitude deviances but with minimal deviances from the original signal otherwise. Control trials and EOG channels were not used in further analyses.

4.1.4 Time-frequency analysis and the phase-locking response

The time-frequency (T-F) representation was derived using the matching pursuit algorithm [178] as implemented by the Matching Pursuit Toolkit for MATLAB [179]. We used a discrete cosine transform dictionary with window lengths of 32, 64, 128, 256, 512, 1024 samples with a window rate of 0.5 Hz. For each trial, we evaluated the top 50 atoms that best represented the section 2.5 seconds before and 1.5 seconds after deviant note onset. The time-frequency figure (Figure 4.2A) shows the sum of coefficients for each time and frequency block over all trials and all subjects, smoothed with a disk filter for better presentation.

To get finer frequency resolution than matching pursuit was able to provide, the neural phase-locking response was computed. For each trial, the 0.6 ms long section post deviant onset was extracted, and all of these sections were concatenated. The power of the Discrete Fourier Transform (DFT) of this signal at 3.33 Hz (1/0.3 s) was divided by the average power at 2.33-4.33 Hz, with the power at 3.33 Hz excluded. While concatenation of segments can introduce an artificial peak in the DFT power, it falls in a different range of the spectrum that does not affect the normalization range. The phase-locking response of the deviants was defined as the average of this normalized value over the top 15 channels where the response was strongest. Channels were allowed to vary between subjects to allow inter-participant variability. The same analysis was performed for the 0.6 s long section immediately preceding each deviant, giving the phase-locking response of the standards. The phase-locking enhancement was found as the difference between the deviant and standard normalized phase-locking responses. The phase-locking enhancement was computed for the following frequencies: 3.33 Hz (target rate), 3.26 Hz, 3.36 Hz, 5 Hz, 10 Hz, 15 Hz, 20 Hz, 30 Hz. Frequencies adjacent to the target rate were chosen with enough distance to exclude spillover of power at 3.33 Hz to nearby frequencies due to lack of sufficient frequency resolution to precisely map power to $3.\overline{3}$ Hz.

To test the effect of different types of feature deviations on the phase-locking response, we performed the same computation steps, but separating the different saliency levels of the tested feature. In the case of pitch deviance, half of deviants

had a low pitch deviance (2 semitone) and half had a high pitch deviance (6 semitone). For the low deviance group, only the deviant and standard sections of these trials were concatenated, and their DFT power was calculated at the same frequencies as listed in the overall case in the preceding paragraph. The deviant power was divided by the standard power to obtain a normalized phase-locking response for the case of low pitch. The phase-locking response for high pitch was found with the same steps performed on the high pitch difference trials. The phase-locking enhancement for pitch was then computed as the difference between the high saliency phase-locking response and the low saliency phase-locking response. The effect of intensity was computed in an identical fashion. The channel selection was allowed to vary between features to account for different sources to perform relevant processing tasks. In both cases, the selection was based on the channels that had the maximum neural response in the high deviance group.

For the overall deviants, as well as the pitch and intensity cases, the phase-locking enhancement was found to be significantly different from 0 only at the rate of the auditory stimulus (3.33 Hz). The rest of the analysis was performed only on this frequency. To investigate the effect of deviance among all feature combinations on the phase-locking response, we computed the normalized response separately for each factorial condition (36 conditions) as described above for pitch: For each condition, only the relevant trial sections were concatenated to provide the phase-locking estimate for those cases. In cases where the concatenated signal length was not long

enough for fine frequency resolution, the closest value to 3.33 was evaluated as the target rate.

4.1.5 ERP analysis

EEG trials were bandpass filtered between 0.7 and 25 Hz and referenced to the nose. Deviant responses were taken as up to 300 ms post deviant onset, and two standards for each trial were extracted as the two notes preceding that trial's deviant: -300 ms to deviant onset and -600 ms to -300 ms from deviant onset. Considering the lack of a silent period between new stimuli and rapid note repetition, the mean voltage of the 50 ms preceding the entire trial served as baseline to the amplitude measurements of the deviant and two standards of that trial. As the initial step to counter the variability of random notes in the background of each trial and possible adaptation effects due to randomized location of the deviant from trial onset, the two standards were averaged to serve as the standard for difference component measurements. Both the standard and deviant that were extracted were subject to influence of notes playing shortly before them, as notes lasted 1.2 s and overlapped every 0.3 s. Further, the background notes had random assignments of timbre and pitch, with slight intensity variance, all of which could lead to variance in the neural response, in addition to the task-irrelevant neural noise already compounded in the EEG signal. The neural noise and varying background notes are random between trial repetitions, but the standard and deviant effect that is sought is consistent over trials.

To isolate these true standard and deviant signals, joint decorrelation (JD) was used. JD is a blind source separation algorithm that finds uncorrelated components from the multi-channel EEG signal that best represent a desired objective, defined here as the maximum trial-to-trial repeatibility as reflected among the full channel space. Once the EEG signals are projected into the optimum JD space, noise elimination was achieved by keeping only the components that preserve 95% of the power in the original signal, then projecting those components back to the original EEG space. This procedure was performed separately for standard and deviant time segments to maximize information relevant to those sections. Finally, the subject difference waveforms were computed as the mean of the standards and baselines subtracted from the mean of the deviants.

First, the difference waveforms at left mastoid, right mastoid, and Fz channels were analyzed. These channels were selected based on the MMN literature, according to which the maximum amplitude of MMN is observed at Fz and amplitude reversal of MMN at the mastoids. Visual inspection confirmed this to be the case. MMN time window was taken as the 40 ms segment of the difference waveform centered at the peak between 100 ms to 200 ms: Negative peak in the case of Fz and positive peak in the case of the mastoids. Significant negative peaks were confirmed for all subjects at Fz by paired t-tests comparing the MMN time window point-by-point to 0, and likewise polarity reversals were observed at the mastoids. Following confirmation of MMN, the trials were re-referenced to the average of the mastoids. Difference wave-

forms were re-computed for all subjects in an identical fashion, for the full 128-channel space. For each channel, the minimum of the grand average difference waveform over subjects between 120 and 180 ms determined the overall peak MMN latency. MMN amplitudes were then computed for subjects as the mean voltage in a 30 ms time window centered at the peak MMN latency. Other ERP component amplitudes were extracted in an identical fashion, centered around peaks with the same width in the following time windows: P1 (positive peak) at 25-75 ms, N1 (negative peak) at 75-120 ms, P3a (positive peak) at 225-275 ms. Due to the rapidly overlapping stimuli, large deviations from trial baseline remained in at the start of each 300 ms section. To counteract this offset, results are presented for P1 and N1 time windows with baseline correction to the average of 0-0.02 ms following each note onset. Results with this baseline did not differ for MMN or P3a, but are presented with the trial baseline, following traditional ERP literature.

Next, the effect of acoustic feature deviations on ERP component amplitudes was analyzed. As each factorial deviance condition repeated only 8 times, and considering the low SNR of the neural data, meaningful ERPs did not emerge for a full factorial analysis. Main effects and two-way interactions were each tested with individual within-subject ANOVAs with Holm-Bonferroni correction for significance. In each case, trials were split into the condition levels and the corresponding number of difference waveforms were formed. For example, to test whether the level of pitch deviance had an effect on the ERP component amplitudes, trials were split into two

depending on whether the deviant note in that trial had a low or high pitch deviance. Two standard and two deviant waveforms, along with two different baselines were calculated, from which ERP component amplitudes were extracted as previously described, and tested with ANOVA.

4.1.6 Inter-trial phase coherence

Coherence of mastoid referenced EEG trials was analyzed on separate frequency bands for each channel individually. The frequency bands were defined as follows: Delta 1-3 Hz, Theta 4-8 Hz, Alpha 9-15 Hz, Beta 16-30 Hz, Gamma 31-100 Hz. Each trial was filtered between these ranges by frequency domain multiplication with the corresponding rectangular filter. The phase of the resulting narrowband signals was computed using the Hilbert transform, defined as the angle of the Hilbert signal. Trials were subsequently segmented into standard, deviant, and post-deviant time sections as follows: Standard as -300 ms to deviant onset, deviant as onset to 300 ms and post-deviant as 300 to 600 ms from deviant onset. All of these time segments are well spaced apart from the onset and offset of epochs, avoiding filter boundary effects. Trial-by-trial coherence (c_{phase}) was computed for each segment separately, as the magnitude of the average instantaneous phase ($\theta(t)$) at each time point t of the segment:

$$c_{phase}(t) = |\Sigma_{n=1}^{N} exp(i * \theta(t))| \tag{4.1}$$

For the overall analysis of whether deviants have higher phase coherence than surrounding standards, all deviants were included regardless of the type of deviance. The effect of acoustic features on phase coherence strength was analyzed by grouping the trials corresponding to different levels of features. For example, to test whether the level of pitch deviance had an effect on phase coherence, trials were split into two depending on whether the deviant note in that trial had a low or high pitch deviance. The deviant time segments were then analyzed for phase coherence. The low number of factorial case repetitions did not allow for a full factorial investigation of coherence. Main effects and two-way interactions were tested with individual within-subject ANOVAs with Holm-Bonferroni correction for significance.

4.1.7 STRF analysis

The cortical activity giving rise to EEG signals was modeled with the spectro-temporal receptive field (STRF). The processing filter acting on the the auditory stimulus $s(t)$ is denoted as $STRF$. The EEG recording $r(t)$ is a result of this processed stimulus signal representation, plus all background cortical activity and noise, denoted as $\epsilon(t)$. The STRF model is then described as:

$$r(t) = \Sigma STRF(f,t)s(f,t-u) + \epsilon(t) \tag{4.2}$$

Estimation of the STRF was performed by boosting [180], implemented by a

simple iterative algorithm that converges to an unbiased estimate. A brief description of the algorithm is as follows. The STRF (size FxT) is initialized to zero, and a small step size is defined as δ. For each time-frequency point in the STRF (every element in the matrix) the STRF is incremented by δ and $-\delta$, giving a pool of $F * T * 2$ possible STRF increments. Among these STRFs, the one that provides the smallest mean-squared error is selected as the increment for the current iteration. This process is repeated until none of the STRFs in the possible increment pool improve the mean-squared error. Here, we further iterate on the STRF by setting the step size to $\delta/2$ and continuing the same process, with 4 step size reductions in total.

For a given channel, the STRFs were estimated for standard and deviant EEG segments separately: Standard section as 1.4 s to 2.4 s from trial onset, and deviant section as -0.1 s to 0.6 s around deviant onset. STRFs were defined for a 300 ms window, reflecting the frequency of new notes. Two-fold cross validation was used to validate STRFs during estimation: Trials were divided into two sections with equal number of factorial repetitions in each section (four repetitions each). The STRF was estimated for one section, and used to obtain an estimated neural response for the other section. The estimated neural response and the actual EEG recording were compared by their correlation. This was repeated for every subject, resulting in $21 * 2$ (number of subjects x number of folds) estimates for the relevant STRF. The STRFs that had a correlation of less than 0.05 were eliminated to remove STRFs with low predictive power, and the remaining STRFs were averaged as the final STRF estimate

for that channel. Using higher fold estimates did not give significantly different results for the overall case with 288 trials. However in subsequent analyses where we divide the amount of trials based on their feature level, the low number of trials resulted in noisy correlation values when higher number of folds were employed, reducing the amount of data used for validation. To keep all analyses consistent, 2-fold validation was used for all cases. Feature STRFs were analyzed by using the deviant segments that correspond to each level of the feature at hand in separate estimations. For example, to test the effect of pitch deviance level, trials were split into two depending on whether the deviant note in that trial had a low or high pitch deviance. The deviant time segments were then used for two separate STRF computations. All STRFs were extracted from the location of strongest MMN, and its 4 surrounding channels. On the Biosemi map, these channels are numbered C21 (Fz), C22, C20, C12, C25. STRF estimates from these channels were averaged to reveal the final STRF, for the overall case and all feature analyses.

All estimated STRFs dominantly show a negative peak and later positive peak at various temporal and spatial locations. To characterize the STRF adaptation effect between standard/deviant, and also different levels of the features Pitch, Intensity, and Timbre-foreground, the elliptical connected shapes around the positive and negative peaks were extracted for each STRF. The maximum amplitude, frequency range and temporal range was computed for these shapes. Significance analysis was performed by permutation testing with 10000 permutations for the overall case, as

well as pitch and intensity increase cases. The permutations were performed on the channel-averaged STRFs, so that if a subject was assigned a label flip, all five channel STRFs of that subject were assigned the opposite label.

4.2 Results

We recorded EEG signals from participants passively listening to a non-melodic train of real instrument notes playing every 300 ms. Compared to traditional oddball ERP studies, the analysis challenges and considerations present in this study are as follows: (1) As there is no silent gap between new notes, and because notes are longer than the repetition rate, the neural recording at any given time is confounded by the effects of previous notes. (2) Neither the "standard" nor "deviant" sounds are repetitions of the same notes, in fact they vary not only in pitch and intensity, but also the instrument. (3) The build-up to the deviant note is of variable length, possibly resulting in differing adaptations between trials. We address the challenges in a variety of ways depending on the analysis, and demonstrate multiple markers of bottom-up attention across the cortex.

Analysis of time-frequency energy around the deviant time revealed a strong response between 3-3.5 Hz in the 500 ms following the deviant (Fig 4.2A). Average power in the 500 ms after deviant onset was significantly higher than in the 500 ms preceding deviant onset between 1-8 Hz (delta and theta bands) (paired t-test,

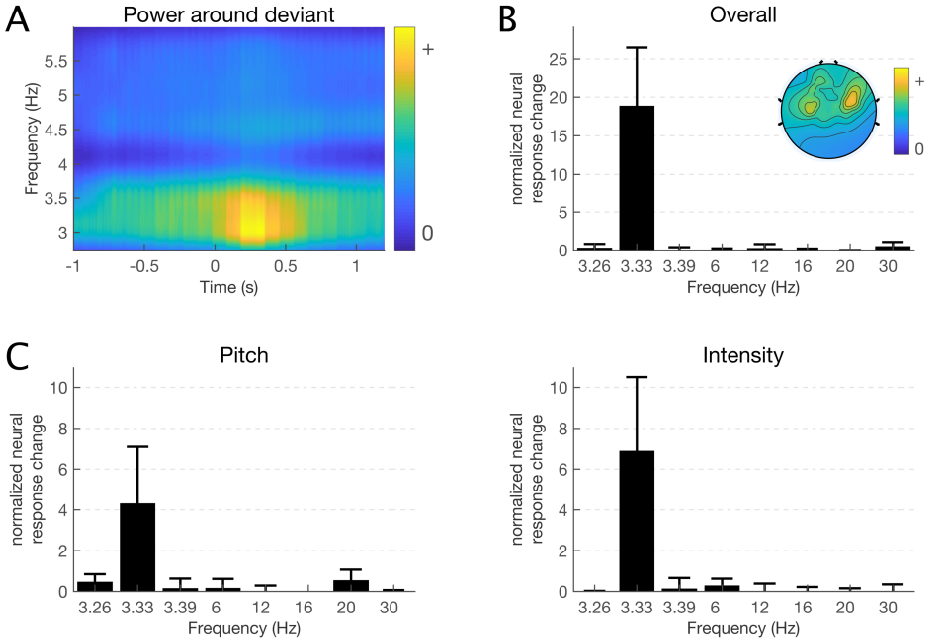

Figure 4.2: Phase-locking of the neural response to auditory stimuli significantly increases for salient sounds. (A) Grand average time-frequency plot shown around the stimulus rhythm (3.33 Hz), where time 0 s indicates the onset of the deviant. A prominent power increase emerges in the 500 ms following salient sound onset. (B) Phase-locking power enhancement for salient notes compared to background notes is calculated at various frequencies. Inset figure shows the topography of the phase-locking increase for an example subject. (C) Phase-locking is modulated by the amount of saliency. Phase-locking power enhancement is shown between high vs. low deviance levels of the plotted feature. As in the overall case, the enhancement is only significant for the stimulus rhythm.

$p < 10^{-6}$). The frequency range of the observed response includes the rhythm of the stimulus, 3.33 Hz. To characterize the effect revealed in the time-frequency plot, we measured the entrainment of the cortical responses to the stimuli before and after the deviant occurrence. The phase-locking response at 3.33 Hz had a significant increase with the presentation of a salient note (paired t-test, $p < 10^{-4}$. Fig 4.2B). The saliency gain on the neural responses was localized to the stimulus rate, no such increase was found for the neural response at close frequencies in the delta and theta

bands, nor in the higher EEG bands.

This enhancement was further computed for the Pitch and Intensity features to examine whether higher deviance among a feature axis elicited stronger phase-locking of the neural signals. This was found to be the case for both features at solely the stimulus rate (paired t-test, $p < 50^{-3}$ for both features. Fig 4.2C). Together, these results demonstrate a prominent saliency gain on the degree of neural phase-locking, and that the gain increases with stronger salience.

Phase-locking power was further investigated for the full factorial design to determine the comprehensive effect of all tested acoustic features and their interactions. The results of a within-subjects ANOVA are given in Table 4.1. The main effects that showed a significant effect included not only Pitch and Intensity but also Timbre-foreground. All interactions that were significant in the behavioral experiment were significant here as well, and 6 new interactions emerged from this EEG experiment that had not been present in the behavioral experiment. Notably, the interaction Intensity * Timbre-background was not significant for any of the behavioral experiments; it emerged as a significant effect for the first time with this measure. The amount of significant effects demonstrated by phase-locked neural responses being much greater than behavioral results for the same stimuli suggests the presence of feature interactions in cortical processes that may not directly translate to cognitive behavior.

ERPs were obtained by averaging the denoised EEG signal in a selected back-

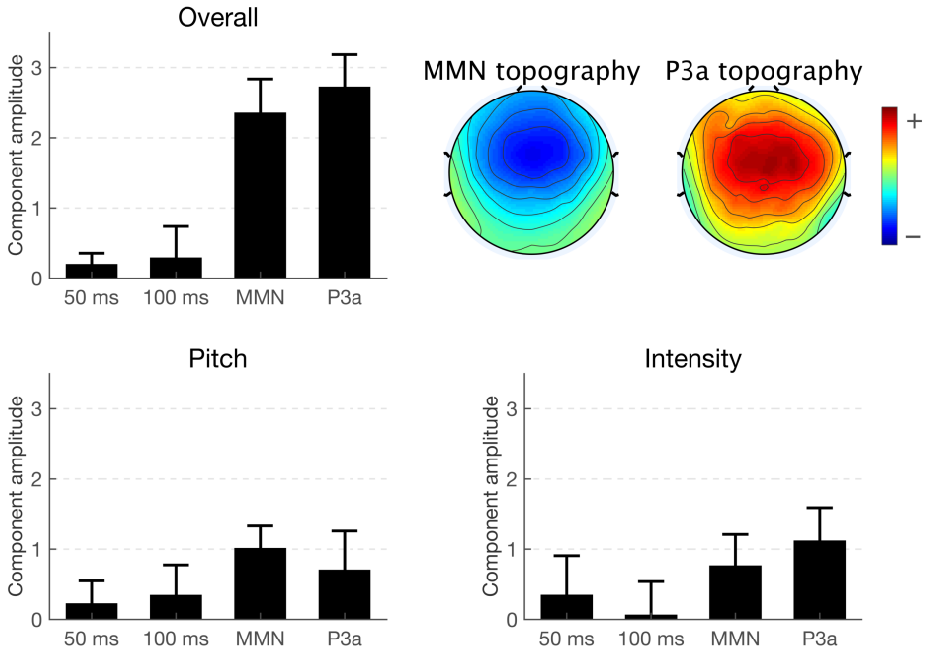

Figure 4.3: Amplitude of ERP components at Fz. In the overall deviant vs. standard case, as well as low vs. high pitch or intensity cases, the MMN and P3a components show a significant effect. No significant amplitude differences were found for P50 or N1 time windows, centered around 50 ms and 100 ms respectively. Similar results are found in surrounding fronto-central channels. Top right panel shows the grand average MMN topography for the overall deviant vs. standard case.

ground note time range as standards and salient note time ranges as deviants. Although a greater SNR is achieved by the averaging of neural signals to repetitions of the same stimulus, we were able to extract ERP components from a pool of deviants that were non-identical, and varied in the features in which they deviated from the background. The common aspect of the deviants is that they violate the regularity in the background formed by standards. The overall difference ERP obtained by subtracting the mean of the standards from the mean of all deviants revealed significant MMN and P3a amplitude effects (paired t-test: $p < 10^{-8}$ for both components at Fz). Both components were further modulated by the increase of deviance among Pitch

or Intensity (paired t-test: $p < 10^{-2}$ for both components and both features at Fz), shown in Fig. 4.3. No significant difference between deviant and standard amplitudes was found for P50 or N1 time windows at any channel. The MMN and P3a components were further analyzed as they are believed to be a marker of automatic deviance detection, and engagement of attention respectively. Amplitudes of both components were found highest in fronto-central electrodes (Fig. 4.3). Latency variances between different feature deviants were not significant. A full factorial analysis of component amplitudes was not possible due to low number of trial repetitions in each factorial case, however, two-way interactions were analyzed at the location of maximum overall MMN and P3a with within-subject ANOVAs (Table 4.1). For MMN, all main effects and the 2 two-way interactions that were found significant in the behavioral experiment were significant (Pitch * Intensity, Timbre-background * Timbre-foreground), as well as Pitch * Timbre-foreground. For P3a, the significant effects were Pitch, Intensity, Timbre-foreground, Pitch * Intensity, Pitch * Timbre-foreground.

Neural entrainment was further investigated in individual EEG frequency bands by quantifying the amount of phase-coherence elicited by salient and background notes. Phase-coherence was overall strongest in the theta band and over central electrodes (Figure 4.4). Salient notes were found to evoke significantly higher phase-coherence than notes in the regular background stream, however, this enhanced alignment did not last after the salient note (Figure 4.4). Background notes that played after the salient note had a comparable amount of phase-coherence to notes preceding the

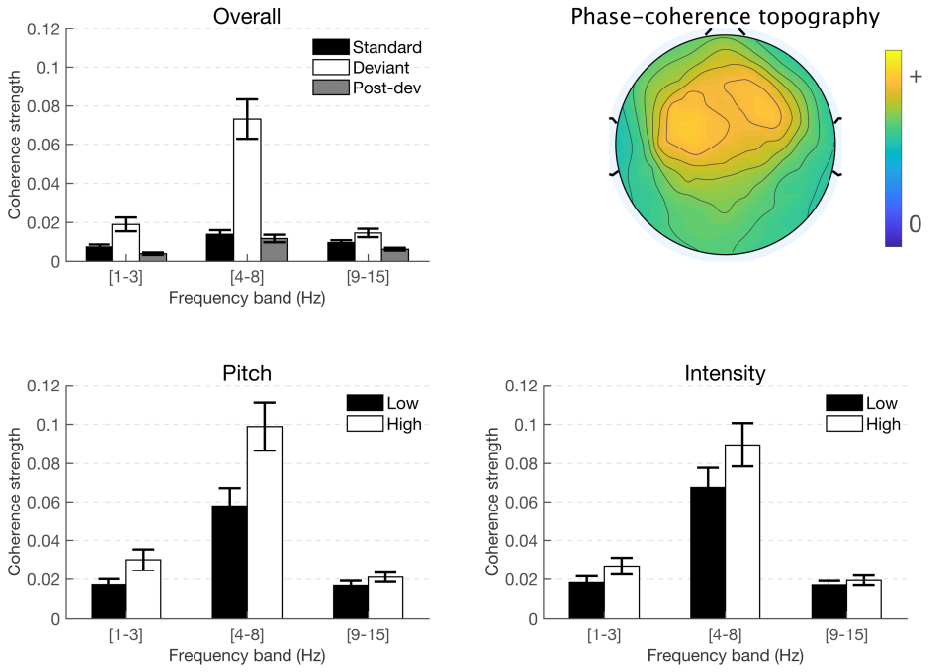

Figure 4.4: Inter-trial phase coherence is stronger for deviants compared to standards, and high saliency compared to low saliency. Although a small effect is seen at the delta band (1-3 Hz), the most prominent effect appears in the theta band (4-8 Hz). The coherence enhancement does not persist after the deviant time window. Top right panel shows the grand average phase coherence for the overall deviant vs. standard case.

salient note. Phase-coherence between notes also increased based on saliency strength. A higher Pitch or Intensity deviance resulted in stronger coherence compared to low pitch or intensity difference (Figure 4.4). Similar to the overall effect, the salience level coherence increase was also observed in the theta band most prominently, with a small amount effect in the delta band, and no effect in higher bands. As with the MMN analysis, the number of trial repetitions was not sufficient to conduct a full factorial investigation of phase-coherence in the theta band; only main effects and two-way interactions were tested. The effects found significant were the same as

Table 4.1: Feature effects on EEG measures of saliency

Effects	F (p)			
	Phase-locking	MMN	P3a	Coherence
Pitch	**12.07 (<E-2)***	**62.43 (<E-6)***	**12.75 (<E-2)***	**30.82 (<E-4)***
Intensity	**20.38 (<E-3)***	**16.58 (<E-3)***	**26.53 (<E-4)***	**24.67 (<E-4)***
Timbre-bg	3.68 (0.03)†	**6.64 (<E-2)†**	1.85 (0.17)†	4.79 (0.01)†
Timbre-fg	**6.92 (<E-2)†**	**5.24 (<E-2)†**	**16.92 (<E-5)†**	**5.82 (<E-2)†**
P, I	**9.57 (<E-2)***	**19.17 (<E-3)***	**5.02 (<0.05)***	**9.72 (<E-2)***
P, T_b	1.29 (0.29)	3.89 (0.03)	1.37 (0.27)	1.26 (0.30)
P, T_f	**5.78 (<E-2)†**	**13.40 (<E-4)†**	**17.77 (<E-5)†**	**28.24 (<E-7)†**
I, T_b	**10.01 (<E-3)**	0.01 (0.99)	1.46 (0.24)	1.62 (0.21)
I, T_f	**8.54 (<E-3)†**	0.79 (0.46)†	2.16 (0.13)†	1.73 (0.19)†
T_b,T_f	**8.84 (<E-5)***	**4.53 (<E-2)***	2.56 (<0.05)*	**6.24 (<E-3)***
P, I, T_b	2.99 (0.06)	-	-	-
P, I, T_f	2.37 (0.11)	-	-	-
P, T_b, T_f	**7.22 (<E-4)***	-	-	-
I, T_b, T_f	**6.39 (<E-3)**	-	-	-
P, I, T_b, T_f	**4.93 (<E-2)**	-	-	-

* Effect was found significant in behavioral experiment (Table 3.1)

† Effect was not found significant in behavioral experiment for this stimuli, but was significant for other stimuli (Table 3.1).

Bolded values indicate significance with Holm-Bonferroni correction for multiple tests.

those for MMN, with the exception of Timbre-background, which was not found to have a significant effect on phase-coherence. The list of significant effects is given in Table 4.1.

All of the measures outlined above (phase-locking, MMN, P3a, phase-coherence) were shown to have elementary dependencies on stimulus saliency: For the overall

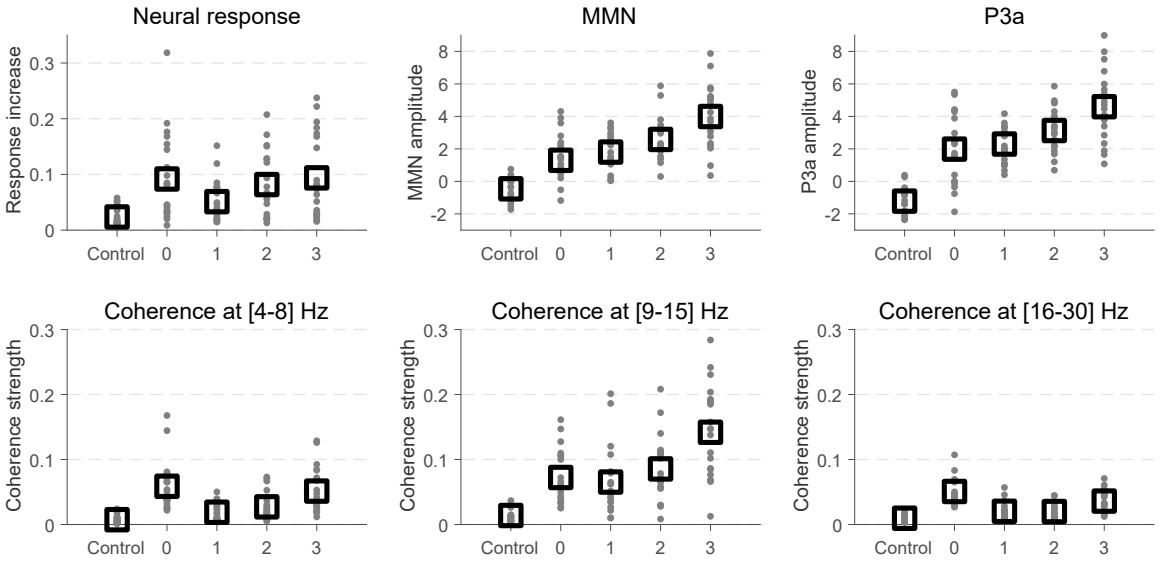

Figure 4.5: Phase-locking response, MMN, P3a, and coherence all increase with greater stimulus saliency. Control indicates the measure for background standard notes. The x-axis denotes the number of features in which the deviant has a high level of difference from the standards. 0 corresponds to deviants with the lowest level of saliency: No difference in timbre, but a 2 db difference in intensity and 2 st difference in pitch. Any change in timbre is counted as high level of salience.

salient note and for increase of saliency among Pitch and Intensity. The comprehensive effect of saliency on the measures was investigated by grouping deviants not based on what specific feature they had changes in, but by how many features they deviated from standard notes, regardless of what feature the deviance was in. The greater the spread of deviance in multi-dimensional feature space, the higher the saliency. An upward trend in strength of all measures was observed as saliency increased. For coherence, the effect was seen most clearly in the theta band. Given that the measures are all modulated by saliency, the influence of each feature was characterized in detail (outlined in Table 4.1). Compared to the behavioral results for this experiment (Table 3.1), two novel effects emerged for every measure: Timbre-foreground and the

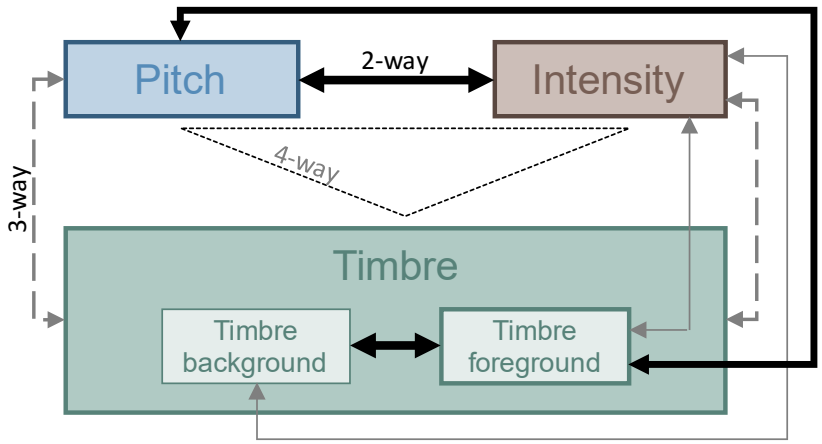

Figure 4.6: Summary of interaction weights based on phase-locking response, MMN, and coherence results as outlined in Table 4.1. Solid lines indicate 2-way, dashed lines 3-way and dotted lines 4-way interactions. Effects that emerged for all measures are shown black, and those that were found for at least one measure are shown grey.

interaction Pitch * Timbre-foreground. It is worth noting that in the behavioral experiment, both of these effects had been found significant for the other two stimulus sets, and it is possible that the reason they had not been significant for this data was due to performance ceiling effects. Two effects that had also been found significant for the other behavioral experiments were Timbre-background and the interaction Intensity * Timbre-foreground. Timbre-background was found to have a significant effect on MMN, but not phase-locking response or coherence, despite having low p-values for both tests (0.03 and 0.01 respectively). One possibility is that there actually is an effect of Timbre-background that did not emerge with stronger power due to noisy measurements, and noise was not fully eliminated in post-processing. Intensity * Timbre-foreground had a significant effect on the phase-locked response, but not MMN or coherence. Consolidated feature and interaction effects are presented in

Figure 4.6.

The dynamic, continuous nature of the experimental stimulus allowed the estimation of neural filters that process sensory input. These filters were modeled after spectro-temporal receptive fields (STRFs) that have been used to characterize the tuning properties of neurons in the auditory cortex. The STRF represents an optimal mapping between the spectro-temporal features of a stimulus and the corresponding neural signal. Unlike time-locked ERP analysis, the STRF is derived as an ongoing convolution filter (the STRF slides over the spectrogram to calculate the response, see Figure 4.7). As a result, negative and positive deflections in the STRF at time t represent not the averaged trend of the neural epochs at time t after onset, but how any unit activation in the stimulus affects the neural response with a delay of t, with no relevance to the concept of an "onset". STRFs can be estimated for the raw unfiltered neural signal, as well as its oscillations at various EEG bands. We find that STRFs estimated for delta and theta bands have a high predictive power, and STRFs estimated for higher EEG bands have low predictive power. As this result agrees with previous work [65, 181] and our phase coherence results, here we report STRFs that were derived from combined delta and theta bands (1-8 Hz).

To estimate STRFs for "standard" and "deviant" sections of the stimulus, separate optimizations were carried out for two different time windows. The deviant time window being longer than the deviant segment results in this window including both deviant and standard sounds. Nonetheless, adaptations driven by the deviant

segment dominate the corresponding STRF, as seen in Figure 4.7A. In this initial analysis, one STRF was estimated per subject at Fz, explaining the response for the standard and deviant segments of all trials as a two-dimensional sliding filter, regardless of the deviance type. Of particular interest is the filter characteristics in areas corresponding to time windows of the neural response that showed significant changes in the evoked ERP, the 120-180ms MMN time window and 220-280ms P3a time window (Figure 4.3). At the corresponding time shifts, both the standard and deviant subject-average STRFs show a response, negative for the MMN time window and positive for the P3a time window. The deviant response is stronger in amplitude and has greater spectral range.

Next, the STRF adaptations between different saliency levels of the tested acoustic features were investigated. In the case of both pitch and intensity, the salience level increase resulted in a response similar to that between the standard and deviant STRFs, where the negative and positive components increased in amplitude and spectro-temporal area. The effect of different foreground timbres was also investigated. Although it is not straightforward to interpret the effect of different instruments, varying spectro-temporal patterns emerged for all three cases, possibly indicating different neural processing for each instrument.

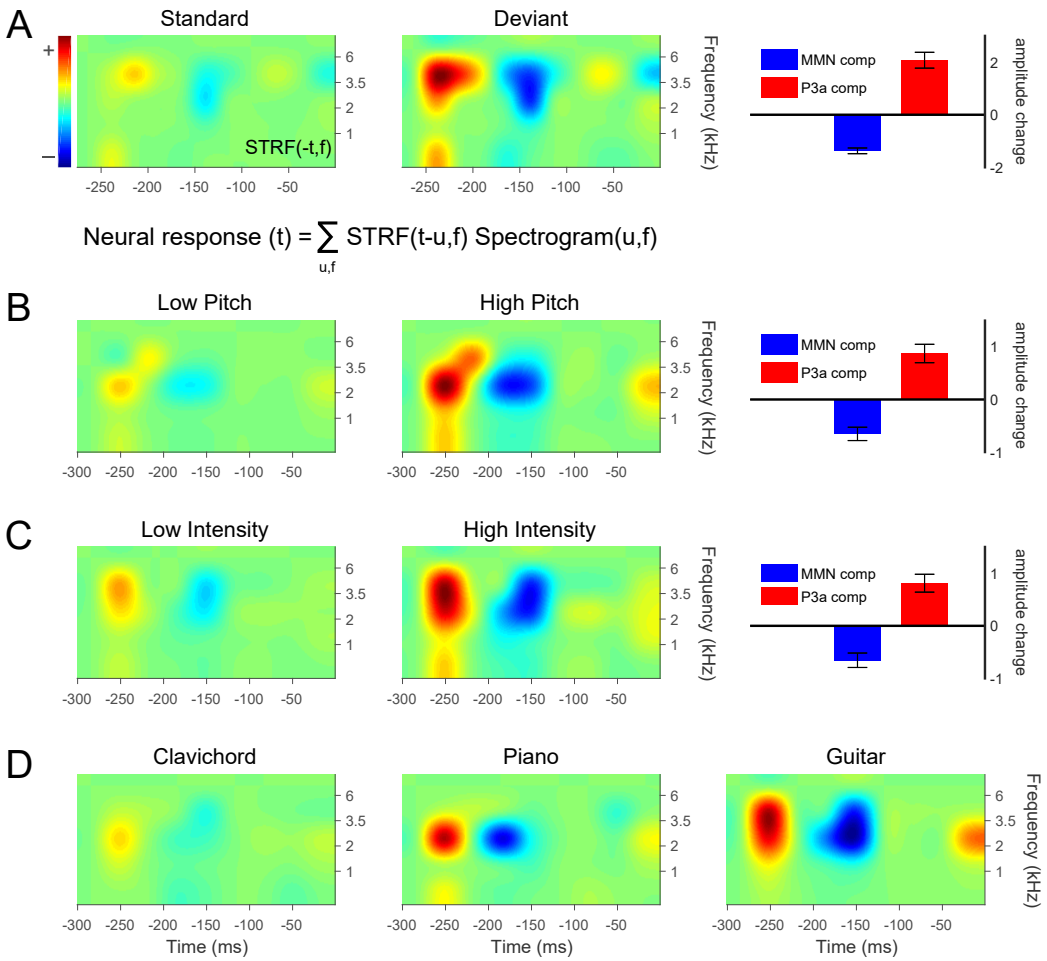

Figure 4.7: Estimated subject-average STRFs. The STRF is found as the two-dimensional sliding filter acting on the stimulus spectrogram that best predicts the recorded neural signal at Fz. (A) STRFs for the overall standard vs deviant case. Rightmost plot shows the amplitude change (deviant minus standard) for the MMN (120-180ms) and P3a (220-280ms) components of the individual subject STRFs. (B-C) STRFs estimated from deviants that belong to specific levels of the tested features. The MMN and P3a components in for features have stronger deflections for higher feature deviance. (D) STRFs estimated for each different foreground timbre. As different instruments cannot be ordered for saliency level, we merely observe that different types of adaptations occur for different feature stimulations.

4.3 Discussion

This study focused on markers of saliency driven bottom-up auditory attention in the cortex. Despite attending to a visual task, neural oscillations entrained to the rhythm of the unattended auditory stimulus, and were modulated by salient events in the auditory modality. This novel finding is complementary to previous studies that have shown neural activity entrainment to the rate or envelope of the attented auditory streams, enhancing the representation of the auditory object under active attention [65, 70, 164, 182]. Our results further demonstrated that higher saliency correlates with stronger phase-locking to the stimulus, particularly in the theta band. The presence of nonlinear feature interactions emerging throughout measures highlights that acoustic features are not processed independently in the auditory pathway. These findings provide new insights into the neural mechanisms processing ambient acoustic scenes, and how salient auditory objects modulate ongoing neural responses to the scenes they are embedded in.

Converging evidence from recent work has characterized the entrainment of cortical oscillations to sensory input as a mechanism enhancing and stabilizing the neural representation of attented objects in the environment [168, 174, 183]. In a close simulation of the "cocktail party problem", given a scene of competing speakers, neural oscillations have been shown to entrain to the attended speech [9, 65, 70, 182]. Of particular importance to the present work is the observation that representations of unattended acoustic objects are nonetheless maintained in early sensory areas [65].

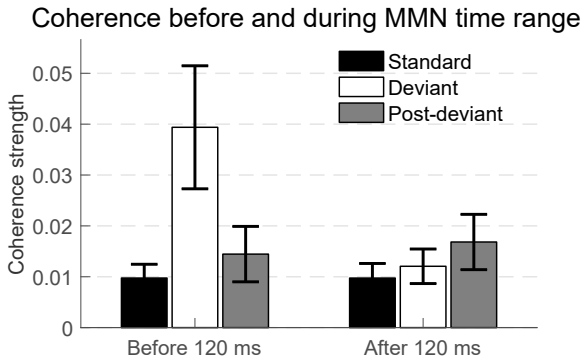

Figure 4.8: Phase-coherence effects are seen before the time range of MMN and P3a, suggesting that they likely reflect different markers of saliency processing in the brain.

Our results complement previous findings, suggesting that the strength of sound representation among the auditory pathway is modulated not only by top-down attention, but also by its salience with respect to the current environmental context. The rhythm of the stimulus falls within the slow modulation range typical for natural sounds such as speech or animal vocalizations, that single-neurons and local field potentials in the auditory cortex are known to phase-lock to [184, 185, 186]. While it is yet unclear whether the observed enhancement in entrainment is a direct result of these single phase-locked neurons, the complex nature of salience in this work, defined on a high-dimensional feature space, make it likely that these modulated responses are the result of cortical circuits involving large neural groups or multiple neural centers.

The presence of a mismatch component followed by an early P3a component in the difference ERP between salient and background sounds provides further support that the entrainment enhancement is associated with bottom-up attention. The fact that our salient sounds deviate among multiple features with respect to a dynamic

background, and a mismatch response is derived from these sounds strongly suggests that the auditory system is collecting statistics about the ongoing environment, forming internal representations of the regularity dynamics in the scene that form the basis to recognize sounds that violate these regularities. Importantly, these regularity violations engage attentional processes as reflected by P3a [187, 188], indicating that the observes effects are not a mere result of acoustic deviance, but attention drawn following a deviance, in a true bottom-up fashion.

Experimental and modeling studies have suggested that MMN and P3a generation is rooted in a hierarchical structure composed of multiple processing centers among the auditory pathway [107] and frontal, parietal and temporal lobes [189, 190]. The specific computations that result in the emergent ERP components in EEG oscillations, however, is still a matter of debate. ERP components, including MMN and P3a, are hypothesized to be a result of either transient bursts of activity across neurons or neural groups time-locked to the stimulus superimposed on "irrelevant" background neural oscillations, or realignment of the phase of ongoing oscillations (phase-resetting) [191, 192]. Previous work has observed MMN responses under increased phase coherence in the theta band with no increase in power coherence, and suggested that MMN is at least partially brought forth by phase-resetting [193, 194, 195, 196, 197]. Our study has presented similar coherence and ERP results; however, a few factors assist in distinguishing these two markers. We noted that the theta phase coherence showed its strongest effect in the time window from salient sound onset to before time

range of observed MMN. Although one possibility of this result is that phase-resetting has triggered a chain of processes leading to the MMN response in cortical areas that are separate than the ones that showed an initial increase in phase coherence, it is equally plausible that MMN and phase-coherence are results of different processes. Regardless of the neural generators, the distinct time windows and varying interactions with different acoustic features that emerged in our work (Table 4.1) point to theta phase-coherence and MMN as being separate measures. If ERP components were generated by additive evoked responses, a related question becomes whether these trial-by-trial amplitude modulations could be the driving factor for the observed phase-locking power enhancement, rather than a true phase alignment of neural oscillations, as the rate of the evoked responses matches the rhythm of the stimulus. We note that ERP amplitude increase is limited to time ranges of the negative and positive components around 150ms and 250ms, and that time-frequency analysis by matching pursuit reveals increased effect of target rhythm on a trial-by-trial basis, making evoked responses an unlikely mechanism for the observed entrainment effects.

Neural mechanisms that generate the recorded EEG signal can be mathematically described as a two-dimensional filter processing the input sound at various frequencies [65, 180, 181]. The key distinction between the ERP and STRF analyses is how each method considers the temporal evolution of neural recordings. The ERP is obtained by time-locked averages of neural signals, thus extracting the positive or negative signal deflections that occur at the same time across epochs. The STRF,

on the other hand, finds a sparse set of filter coefficients that best explain every instance in the epoch as a function of the past 300 ms of input sound. The time windows of significant ERP components (MMN and P3a) in the STRF show areas of inhibition and excitation, respectively (Figure 4.7), that become stronger to account for higher saliency. The EEG STRF for the musical stimulus used in this study has significant predictive power at only 1-8 Hz, indicating that the neural signal encodes slow temporal dynamics of the acoustic input most prominently.

It could still be argued that this study is not free from effects of top-down attention, because while the experiment subjects are given movies to watch, their attentional state is not explicitly controlled away from the auditory domain. Even if attention was fully controlled to remain in the visual domain, task demands are likely to influence the perceived saliency of ignored sounds, requiring a sound object to be more salient to attract attention. These confounds can be elucidated with future experiments that study the strength of auditory event locked entrainment under a varying level of task difficulty. Although the current study paves the way for studying pure bottom-up auditory attention, the challenges involved with interpreting the meaning of auditory saliency without introducing tasks in other domains reflects the more complicated nature of salience across modalities other than vision, where saliency maps are hypothesized to be encoded among the visual pathway [27]. Nonetheless, the entrainment measures employed in this study can be used for natural scenes to decode salience responses from EEG or MEG recordings, possibly eventually

producing a ground-truth saliency dataset for the auditory domain as an analog to eye-tracking data in vision.

The increasing amount of results on the effect of attention on cortical entrainment in recent years has spawned a number of hypotheses about its underlying neural processes, and although ERPs and the MMN component is well characterized, its underlying neural processes also remain uncertain. Computational models are likely to be helpful in testing the extent to which each hypothesis can explain experimental results. While several distinct auditory saliency models have already been proposed, the lack of a standardized and intuitive dataset describing human bottom-up auditory attention is one of the primary reasons that modeling efforts in the literature remain sparse [151]. Neural recordings, rather than behavioral measures, are likely to lead to an objective automatic measure of natural stimulus saliency, and we suggest the study of cortical entrainment as a target for saliency research. Finally, we stress the importance of feature interactions in the auditory domain, emerging from previous behavioral results, and across all EEG measures in the present study. Few auditory attention models consider the interaction of features, instead processing them in parallel and aggregating features with equal weight. It is possible for the addition of feature interactions to raise the performance of a "bad" model to be better than a "good" model, thus caution should be exercised when interpreting models that do not incorporate this aspect of audition that is so prevalent among all saliency measures, whether behavioral or neural.

Chapter 5

Investigating selective auditory attention

5.1 Introduction

In a cocktail party, attention can be directed by multiple external and internal factors. We can follow a speaker while ignoring background conversations, search for the source of a high-pitched laugh or a buzzing coming from a particular direction, and involuntarily be distracted by a loud crash. The psychological and neural mechanisms that allow us to efficiently locate and track perceptual units of interest in an environment with competing sources are still a matter of debate, with much of the conversation centered in the visual domain. A considerable amount of evidence suggests that visual attention can bias processing in favor of spatial locations, stimulus

features such as color or motion, or whole objects. Feature-based attention is primarily observed as enhanced activation of cortical populations that process the attended feature [83, 198, 199]. The key distinction between feature-based and object-based attention lies in the processing of unaattended features: Object-based attention theories posit that attending to a feature of an object will cause the whole object to be selected, leading to enhanced processing of the unattended object features even when these other features are irrelevant to the task [200, 201, 202]. However, recent work has shown that selection based on features can be associated with active suppression of unattended features in the scene, including the task-irrelevant features of selected objects [203, 204, 205]. To reconcile these seemingly contradictory results, current theories of visual processing suggest that feature- and object-based attentional mechanisms coexist in the visual pathway [206, 207, 208, 209].

Selective attention to sound has been shown to modulate responses in the auditory cortex, using single neuron recordings [210, 211], electroencephalography [61, 212], magnetoencephalography [149], positron emission tomography [63], functional MRI [213, 214], and electrocorticography [215]. Early studies investigating attention to auditory space and frequency failed to find feature-specific enhancement [60, 63, 216], reasoning that auditory attention acts not on low-level features, but on integrated object representations [14, 21, 63], particularly in secondary auditory cortical areas [60, 217]. A considerable amount of imaging work has since gathered evidence in support of feature-based attention in primary and non-primary auditory cortices [64,

218, 219, 220, 221, 222, 223, 224], though support for suppression of unattended features as part of feature-based auditory attention is yet lacking [221]. It is not clear whether the lack of a difference in feature effects observed in some of these studies [14, 60, 63, 217] is due to co-selection of unattended features as in object-based visual attention. Instead, support for attention acting on auditory objects came from behavioral studies demonstrating that attention is affected by continuity in both task-relevant and task-irrelevant features of attended objects [225, 226]. These results are complemented by imaging studies demonstrating enhanced cortical representations of attended speech (the "object") in a multi-speaker paradigm [9, 65, 182, 227].

While the studies outlined above demonstrate that attention can enhance feature and object representations in auditory cortices, it is unclear whether the observations are a product of a single attentional mechanism, or whether feature-based and object-based auditory attention differ. One possibility is that attention acts on auditory objects, and results from feature-based studies can be interpreted as enhancement of the attended object's features. To address this possibility, we tested the effect of these two types of attention in a single experiment, using the same stimuli for all attentional conditions. Listeners performed an engaging "cocktail-party" task, with concurrent male and female German speakers that occasionally changed direction between left and right (Figure 5.1). To contrast the two types of attention in this highly challenging scene, attention of subjects was directed to the male speech (object), or all speech coming from the right direction (feature) in separate blocks. All speakers had a similar

pitch range, and the task was to detect parts of the speech that were manipulated to have abnormally high or low pitches, outside the pitch range of the regular scene. We hypothesized that detection would improve if attention was directed to relevant streams. A free-listening block (global attention) was performed first, serving as a baseline for detection of salient events in the scene. Our results suggest that feature-based and object-based attention do not reflect the same process, and further are nonlinearly affected by saliency.

5.2 Methods

5.2.1 Stimulus design

Experiment stimuli consisted of simultaneously played German audiobook narration extracts by three speakers (a young female, an old female, and a young male, all adults) that were chosen to have a similar vocal pitch range. Each speaker started at one specific direction (right/left/center) at the start of every trial. After a few seconds, one or more speakers began to move towards the opposite direction. Upon reaching the opposite direction, they remained there for a few seconds. This moving pattern happened 1-3 times. The speaker movements were constructed such that at every point in time, there would be no gaps of speech at absolute right or absolute left, so as not to bias the audio towards any direction. Further, speakers never overlapped in any direction, except for the brief moments when one speaker was reaching

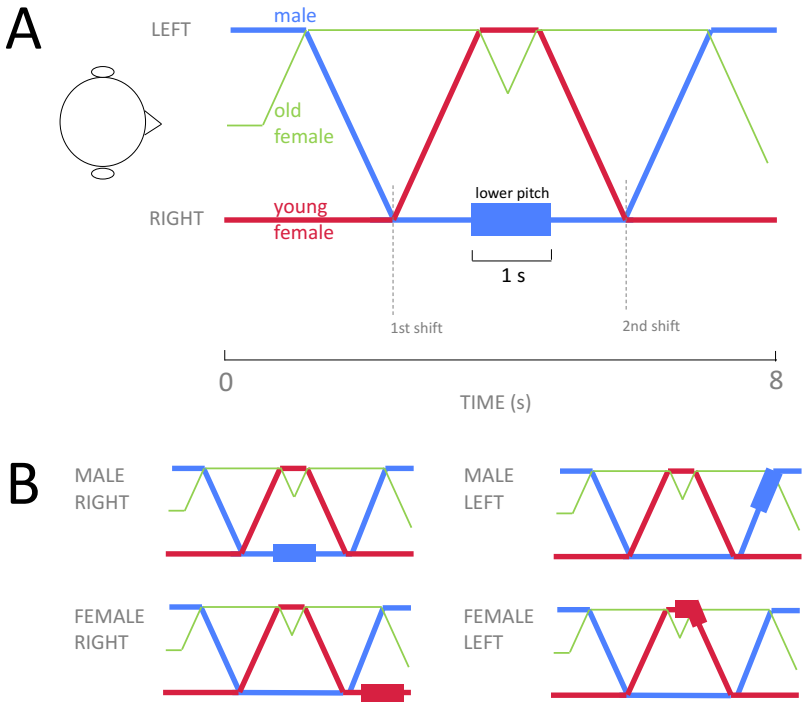

Figure 5.1: Setup of the cocktail-party stimulus used in this experiment. (A) Three German speakers narrating book sections start at left/center/right in each trial, with subsequent occasional shifts in direction. Targets are 1 s long segments of the natural speech manipulated to have a pitch outside the pitch range of the regular speech (3 st difference, lower for male targets, higher for female targets). (B) Four conditions are tested for the target placement, crossing the male and young female with left and right. To realize the feature attention task as attending to right, right targets are fully in the right direction. Left targets are allowed to vary between center and left to give the global perception that targets can appear anywhere in space. Shown in the figure are some example possibilities for left target spatial placement.

a direction just as another speaker was leaving it.

The male always started from the left, younger female from the right, and older female from the center. The right stream was restricted to have only the male or young female. The older female's first direction change took her from center to the left, and her direction changes happened between center and left. No speaker lingered in the center after stimulus onset. Speakers were either moving between directions,

or at right/left ends.

Stimulus length varied between approximately 5-10 s. Speed of moving between directions was constant, 1.2 s from one side to the other. Direction shifts started at a random time after the first second, the number of shifts was the maximum possible number the trial length allowed. Timing of direction changes were randomly selected as long as the stimulus constraints were met.

5.2.1.1 Scene parameters

Speech segments used for all speakers were manually extracted from public domain Librivox German book narrations (Male: https://librivox.org/ein-vade-mecum-fur-den-hrn-sam-gotth-lange-pastor-in-laublingen-by-gotthold-ephraim-lessing/, Young female: https://librivox.org/menschen-im-krieg-by-andreas-latzko/, Old female: https://librivox.org/das-letzte-maerchen-by-paul-keller/) recorded at 22050 Hz. Eighty-one male segments were extracted from chapters 2 and 5, 61 young female segments from chapter 1, 58 old female segments from chapter 10. The overall pitch range was approximately A2-D4. Segments were chosen to have prosody to sound like spoken single sentences, with no regard to meaning of the words spoken or whether the segment contains actual sentences; thus segment length was determined primarily by the dynamics of the speech, each approximately 6-13 s before processing. As subjects would need continuous speech to follow speakers in a very busy scene, segments were manually processed to remove silent periods, including narrator pauses and words

spoken very quietly. Fifty trials were constructed by selecting one unique segment for each narrator such that the length difference between the three segments would be less than 300 ms. All three blocks (global, feature, object) used the same 50 trials. Trials were 5.1-9.6 s long, and were assigned to experiment conditions randomly.

All stimulus construction and experiment analysis were performed with Matlab software. Experiments, including the interface and sound delivery, were performed in Presentation software.

5.2.1.2 Spatial parameters

The three speakers were positioned in simulated 3-D space with a head-related transfer function (HRTF) recorded on a mannequin (Neumann KU 100) under the same conditions that human HRTFs are recorded in. The NH172 HRTF was used from the ARI HRTF database. Trajectories for each speaker were constructed between -90 (left) and 90 degrees (right) denoting their position for each time point. Spatial dynamics of the scenes were increased by adding jitter to the trajectory when speakers had stable position at left, center or right: Instead of a straight trajectory, a sinusoid with a period of 5 spanning 50 degrees was inserted. At direction change times, the trajectory moved from the middle of the left/right jitter (-65 and 65 degrees) in a linear line lasting 1.2 s. The old female speaker was the only one who moved between left and center. Movements for this speaker took only as long as necessary to fill in gaps in the absolute left, but with the same speed as movements of the other speakers.

5.2.1.3 Target construction

Targets were 1 s long segments of the ongoing speech that were manipulated to have a modified pitch. Right target times were selected randomly after the first second, as long as the speaker trajectory was in the right side for the entire duration of the target. Left targets were allowed to have a trajectory that varied between center (0 degrees) and left. This allowed for global perception that targets could appear almost anywhere in space, but have a separate right-only stream to test feature attention. Male targets never appeared before the first direction change of the male from left to right. Young female targets could appear before the first direction change, in the right side. There were no old female targets.

Pitch manipulation was performed by time dilation (for male) and compression (for female) with a phase vocoder [228]. To move the segment pitch out of the range of the pitches occurring in the scene, the male targets had lower pitch, and female targets had higher pitch, with a difference of approximately 3 semitones in either direction. Target segment onsets and offsets, as well as speech immediately prior to and following the target were smoothed by 10 ms long ramps to avoid abrupt transitions in sound.

5.2.2 Experiment procedures and participants

The task was to detect a segment with an unusually different pitch. Fifty trials with unique sentences were constructed, 10 of which contained no pitch altered segments (control trials). The target specification in the remaining trials were as follows: 10 male-right, 15 male-left, 10 female-right, 5 female-left. The target was a 1 s long segment of the speech that was altered to have slightly higher (if female) or lower (if male) pitch. The right target segments were in the absolute right for the duration of the target, but the left targets could be anywhere between center and absolute left. Stimuli were delivered with headphones (Sennheiser HD595).

The experiment had three blocks. In the first block (global), subjects were presented the stimuli and asked whether they heard the target segment. In the second block (feature), subjects were instructed to pay attention to speech on the right ear and ignore the left ear to the best of their abilities. In the final block (object), subjects were instructed to pay attention to the male and ignore both females. In the feature and object blocks, subjects similarly reported targets, including those they heard that were not in the instructed stream (distractors). Subjects were randomly assigned to perform the feature or object block before the other, but the global block was always performed first. The same 50 trials were presented in all three blocks, with trial order randomized in each block. A training section before the experiment assured that subjects understood the global task. Subjects had no prior knowledge of the feature/object tasks until the end of the global block.

A total of 78 subjects (43 female, aged 18-31 years) participated in the experiments after giving informed consent. Sixteen subjects total were removed from analysis due to being unable to perform the feature or object tasks, determined by whether their hit rate of distractor trials were higher than of target trials (e.g. hitting female deviants more than male deviants in the object task). Thirty-nine subjects participated in the initial version of the experiment. However, the density of target times in the trials was too sparse to construct meaningful temporal analyses. Two new sets of 50 trials were created with the same sentences and same target design, the only difference being a greater variance of target time. Twenty-three subjects performed the experiment with one of these two sets, assigned randomly (11 and 12 each). All stimulus sets produced similar average hit, control, and distraction rates, so they were grouped together for all reported results.

5.2.3 Saliency classification of trials

Trials including targets were analyzed with the saliency model in [38]. The model builds statistical predictions among a variety of acoustic features and derives saliency among each feature as a function of deviance from the predicted feature value at each time. Saliency values are boosted depending on the saliency at other features. To reduce noise, here we only calculated interactions for the maximum spikes along each feature at the duration of the target, resulting in only one feature vector per trial. To classify saliency level instead of saliency existence, subject responses were used as

ground-truth data. The trial was assigned 0 if less than half of the subjects heard the target, 1 if half or more subjects heard the target. Finally, each trial was classified as low or high saliency depending on whether saliency predictions from logistic regression were less than 0.5 or greater than or equal to 0.5 respectively.

5.3 Results

Performance between the three attention conditions shows a clear ordinal pattern, with global attention resulting in the lowest performance, and object attention the highest performance (Figure 5.2). This pattern is reflected both in increased hit rates, and decreased false responses (Pairwise t-tests for hit rate: global x feature: $p < 0.001$, global x object: $p < 0.001$, feature x object: $p < 0.002$). False responses to control trials, which have no target, decrease prominently with directed attention, with no difference between feature and object cases (Pairwise t-tests for control-false rate: global x feature: $p < 0.001$, global x object: $p < 0.001$, feature x object: $p = 0.04$). However, another source of false responses in this experiment is distraction trials in attention directed tasks. Distraction trials are not part of the stream that should be attended to, but contain targets in the opposite stream. Distraction rate is significantly higher for the feature task compared to the object task (Pairwise t-test for distraction-false rate: feature x object: $p < 0.001$). These results suggest that object-based attention is stronger than feature-based attention. Results are found to

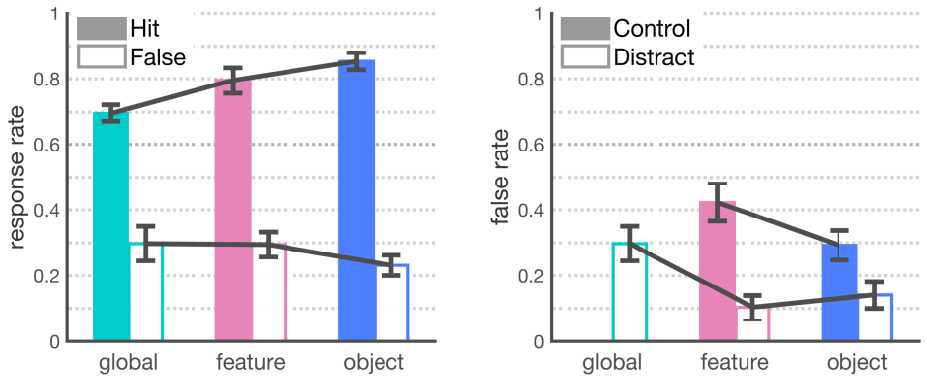

Figure 5.2: Target detection performance for different types of attention tested in this experiment. Global attention represents free-listening, providing a baseline to evaluate selective attention. Directing attention to an acoustic feature (right side) and object (male speech) results in progressively increased hit rates and decreased false rates.

be the same regardless of whether the feature or object task was performed first.

An interesting effect on hit rate emerges when saliency of targets is considered. Grouping targets by their relative saliency, computed with a bottom-up attention model [38], reveals that selective attention strongly boosts perception of low-salience targets that would otherwise be missed, whereas high-salience targets are detected to a similar degree under global or selective attention (Figure 5.3). Notably, only 60% of low-salience targets are detected without directed attention, barely above chance level (50%). Both feature and object tasks significantly raise performance, with object-based attention resulting in higher detection than feature-based (t-test, $p < 0.05$). Interestingly, the advantage of object-based attention over feature-based attention disappears for high-salience targets. As a result, both modes of selective attention show a significant nonlinear interaction with saliency ($F_{feature} = 5.13$, $p < 0.05$, $F_{object} = 14.54$, $p < 0.001$).

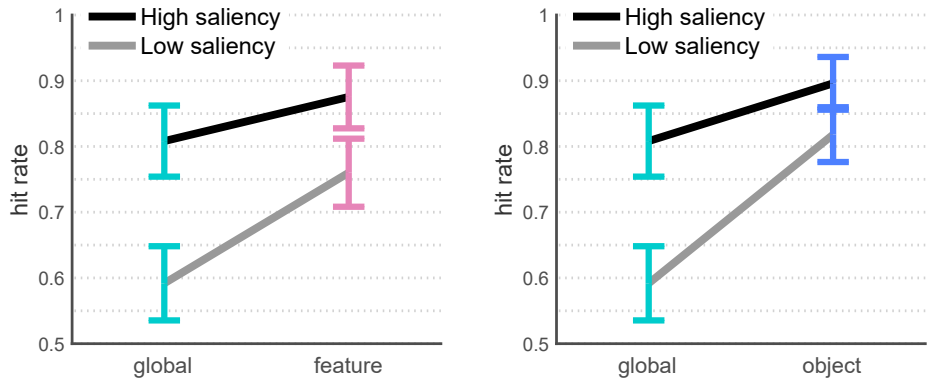

Figure 5.3: Selective attention interacts with saliency. For both feature-based and object-based attention, performance is differentially modulated by the salience level of the target. Specifically, detection of low-salience targets is significantly boosted by both feature- and object-based attention, and a comparatively minor boost is seen for high-salience targets.

An examination of the time course of target detection after attention reorientation as a result of speakers changing direction reveals differing amounts of build-up for the tasks performed (Figure 5.4A). Surprisingly, the build-up effect is seen only for targets that happened before the first shift, or following the first shift, the "early" shifts (see Figure 5.1 for an illustration of shifts in trials). Targets that happened after the second or third shift, the "late" shifts, do not show this pattern, instead all three attention tasks show a stable hit rate level over time (Figure 5.4A, right). In the early shift case, the build-up lasts roughly until 1.2 s after shift before stabilizing.

The overall build-up pattern shown in Figure 5.4A is reflected in both low and high saliency trials. Figure 5.4C shows that the build-up angles are most prominent for low saliency targets that happened after early shifts, but for high saliency targets, there is less build-up. Particularly, detection of high saliency targets is not affected by shifts under global attention. For feature and object attention, however, there

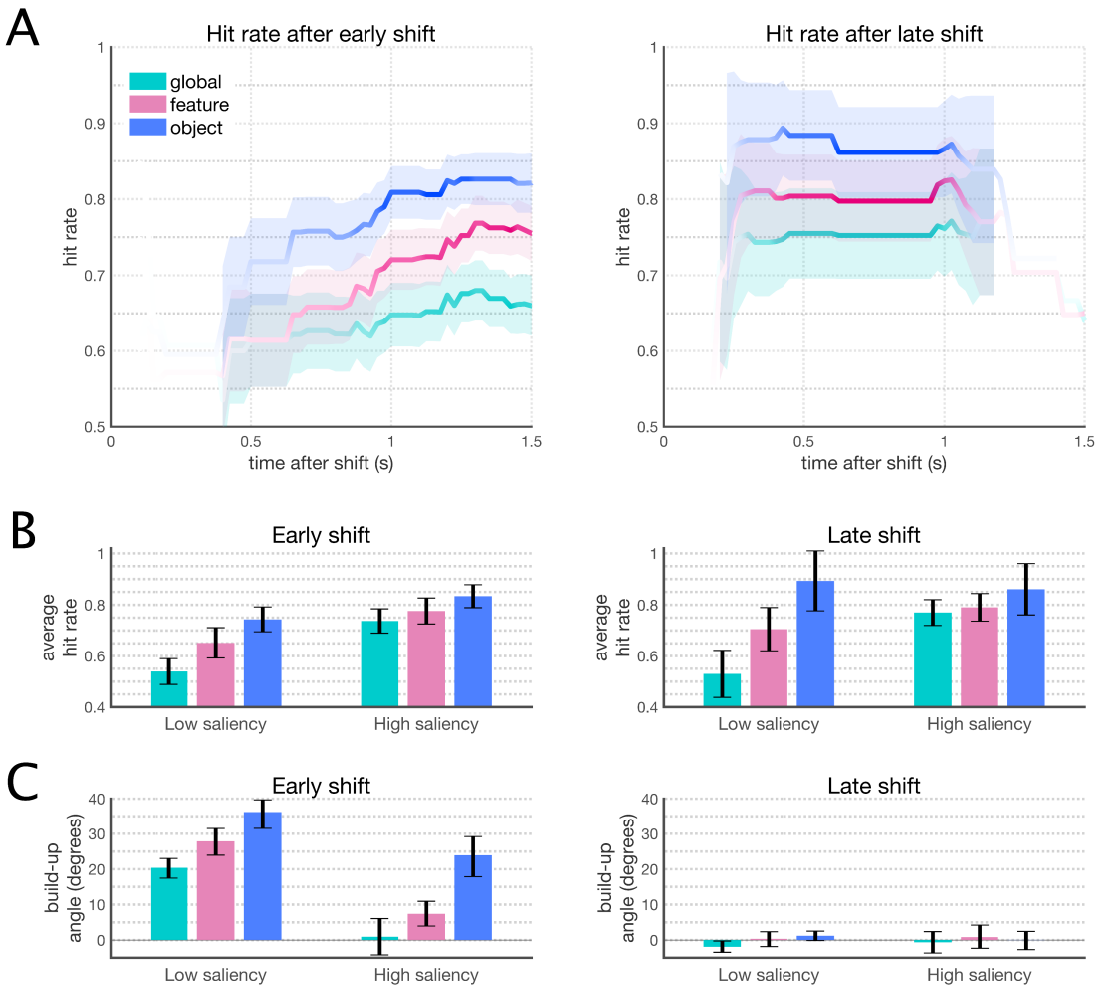

Figure 5.4: Temporal build-up of target detection under different types of attention. (A) Time course of hit rate for early (before or after the first speaker change) or late (after the second or third speaker change) attention reorientation. A shift denotes a change of speaker between the male and a female. Rates at every time t denote the average hit rate of targets that are playing at time t: Targets that started between $t - 0.8$ s and t s. (B) Average hit rates analyzed by shift time and saliency level. (C) Build-up angle of the time-course for different shift times and saliency levels. Angle is found by fitting a line at the build-up time window.

is still some build-up. Interestingly, there is no build-up observed for low saliency targets that happened after late shifts.

The time-course of target detection also demonstrates that the ordinal pattern

between the three types of attention are reflected throughout the entire time course of the scene. The advantage of directed attention is the smallest for high saliency targets, whether they happened early or late in the trial (Figure 5.4B). Low saliency targets show the biggest sensitivity to the direction of attention when they came after late shifts, suggesting a refinement of selective attention over time. These results are supported by a statistical comparison between task, saliency, and shift, revealing a significant interaction between saliency and shift ($F = 8.15$, $p < 0.01$), and saliency and task ($F = 4.65$, $p = 0.013$). All main effects are significant ($F_{task} = 9.1$, $p < 0.001$. $F_{saliency} = 58.35$, $p < 0.001$. $F_{shift} = 15.6$, $p < 0.001$), but no significant three-way interaction or interaction between shift and task is observed.

5.4 Discussion

The current results demonstrate that auditory attention can operate in at least three unique ways in continuous, natural sound environments. We find that selective attention can be directed in a feature-based and object-based manner, and that both present a clear advantage over free-listening, global attention. The latter contrast is of particular significance to explicitly illustrate that effects observed in directed-attention tasks are not caused only by the inherent dynamics and acoustics of the presented scenes, and to establish a baseline from which the degree of attentional enhancement can be quantified.

Object-based attention appears to be stronger than feature-based attention, resulting in enhanced perception of small acoustic deviances that would otherwise go unnoticed in a busy scene. This result is especially powerful in our paradigm manipulating different types of attention on the same set of stimuli. Enhanced sensitivity to events of interest in the "cocktail-party" is complemented by decreased distraction by background events in object-based attention. These differences support the hypothesis that selective attention can operate at different levels in a hierarchical framework of object formation based on the binding of features into proto-objects and objects, with feature-based attention actively biasing early acoustic representations. Although ultimately the unit of perception is an object (a speaker in our experiment), we note that even after performance builds up and stabilizes, feature-based attention remains weaker than object-based attention (Figure 5.4), suggesting different underlying processes rather than simply finding the target object that has the desired feature and defaulting to object-based attention.

Without task goals in global attention, perception is largely driven by acoustic salience. When the target is not very salient, a dramatic enhancement in perception is evident for both feature- and object-based attention from the global baseline, with object-based attention resulting in the best performance. Interestingly, much of the advantage of directed attention disappears when targets are already inherently conspicuous. While directed attention still has a small advantage over global attention, the effect does not appear to depend on whether the direction is feature-based or

object-based. These results imply that top-down selective attention has little effect when bottom-up attention is highly informative. These results further demonstrate that the distinction between feature-based and object-based attention is most clearly observed under high task demand in a busy, natural scene, and may not be readily apparent in scenes with few discrete stimuli with less competition for attention.

A fundamental part of our paradigm is the frequent direction change of speakers within trials. We observe a decline in performance followed by rapid reorientation of attention for all tasks, but this effect is present only for early speaker shifts. As spatial configuration changes become part of the regularity in the scene, attention does not suffer following shifts and reorientation is not apparent. This seems to be the case even for hard to detect, low saliency targets. On the other hand, low saliency targets show the fastest build-up in reorientation after early shifts, with object-based attention building up most rapidly. Overall, global attention is least affected by feature changes in the scene, further reflecting that it is primarily driven by target saliency, irrespective of object parameters. It is important to note that the build-up effects observed, rapidly stabilizing in less than 2 seconds, likely represent a different effect than object formation in a scene, known to evolve over seconds [38, 229, 230]. The direction change of speakers does not reset attention entirely, instead reconfigures the existing regularities in the scene, thus representing object change rather than object formation. That objects and feature statistics are refined over time is instead evident in the performance level differences in the early and late

sections of the scene: Both object-based and feature-based attention result in higher performance after late shifts, but only for low saliency targets. Highly salient targets and global attention nullify the temporal build-up, although it is possible that had we tested targets near the start of trials, at less than 1 s, we might have seen a global build-up effect, as in [38].

Although in this work we have only tested feature-based attention to acoustic location, neuroimaging results demonstrating attentional modulation for a variety of features suggest that similar effects could possibly be seen for directed attention to other features. This is especially the case considering that there does not appear to be feature-dependent differences in feature-based attention in vision, and that visual and auditory attention seem to share many common mechanisms [231, 232]. One distinction from vision, however, is in the treatment of the spatial dimension. Space-based attention has traditionally been treated as a separate form of selective attention in vision [206, 233], though it has been suggested that it could be unified under the same framework as feature-based attention [234]. However, there is little evidence supporting space as a special feature in audition. Studies that have investigated the effect of attention to frequency and space have suggested the two features operate under the same fundamental process [218, 226]. It is also worth considering that space in audition is derived from neural computations on signals reaching the two ears, in a similar manner to pitch or other acoustic features. Even if spatial attention differed significantly from feature-based attention, the current results still demonstrate that

auditory attention can operate in three distinct mechanisms, with spatial or feature-based attention differing from global and object-based attention.

An important factor to consider when evaluating feature-based attention is the effect of inter-trial priming. It has been argued that perception in the lab is highly biased by attentional demands of the previous trial [235], introducing a bottom-up mandatory enhancement that might be sufficient to explain increased performance in detecting the tested feature. Crucially, evidence demonstrating that visual feature-based attention cannot override previous trial priming [236, 237, 238, 239] puts its classification as a top-down, volitional type of attention into question. Recent auditory attention studies have varied attentional direction to pitch and location between trials, establishing top-down prepatory neural responses to either feature without priming influence [14, 240, 241]. Although the current study has used a block design with possible trial-to-trial priming effects, our results support the claim that the observed feature-based effects are largely caused by active selective attention. Despite progressive enhancement in perception within trials as illustrated in Fig 5.4, the build-up reset at every trial for all forms of attention, whereas feature priming between trials would suggest that the effect of build-up would carry over to the next trial. That our stimulus presents a challenging cocktail-party scenario likely plays an important role in forcing volitional attention to perform the task. Indeed, it has been observed that the effect of priming is reduced when target search becomes difficult [239]. These points highlight the importance of using complex, natural scenes

in probing attention.

Our results unify evidence from pyschoacoustical and imaging studies by demonstrating that auditory attention can narrow its global focus in both a feature-based and an object-based manner, further illustrating that the distinguishing characteristics between modes of selective attention are most prominent for low-saliency targets. It remains to be seen whether the different types of attention represent parts of the same neural mechanism, and to what extent they interact to drive perception and behavior in natural settings.

Chapter 6

Conclusion

This dissertation has provided a multifaceted approach to characterizing auditory attention. Through behavioral and neural experimentation, as well as computational modeling, we have demonstrated that auditory attention is modulated by pitch, loudness, and timbre contrast with respect to the regularly occurring sounds in the scene. This dependence is reflected in behavioral measures, and throughout markers of attention obtained from EEG recordings, as described in Chapter 4. Specifically, bottom-up attention can be observed in EEG recordings in the strength of entrainment to the acoustic stimulus, phase-coherence in the theta band, and the combination of MMN and P3a responses. All of the aforementioned behavioral and EEG responses are modulated not only by single features, but are also modulated nonlinearly by combinations of features. In other words, different combinations of pitch and loudness difference have different perceived salience. This finding is of par-

ticular significance as nearly all auditory saliency models in the literature, following in the visual modality's footsteps, treat feature dimensions as parallel, only combining saliency computed independently among different features at the end. We further corroborated the importance of feature interactions when we used the attention model in Chapter 5 to derive low or high saliency levels based on subject detection level in the experiment, using an entirely different dataset that was significantly more complex than the stimulus used in previous experiments. Without feature interactions, the model is not able to come close to matching human performance. To put the impact of interactions into perspective, consider that adding this step to two models can reverse their effectiveness in matching human performance or ground-truth.

We have presented a computational attention model that maps incoming sound to a high-dimensional feature space covering pitch, loudness, time-frequency evolution, frequency modulations, and temporal modulations, with saliency found among each feature. The crux of the saliency computation lies in predictive coding, tying into a recent global hypothesis of brain function that propose the brain as a system that constantly makes predictions and compares the events that occur to those predictions to drive behavior [40, 242]. The model learns statistics of patterns that regularly appear among each feature, and finds salient events as those that do not match any of the repeating patterns in any feature. As discussed above, feature interactions play an important role in consolidating saliency estimates along individual features into a global saliency prediction in time. Although we presented a salient/not-salient

tailored optimization with the model specification in Chapter 3, we also performed a slightly different type of optimization for low/high level of salience for a much more noisy dataset in Chapter 5. Further, although not described in this dissertation, we have used the model as an abnormality detector for lung sounds, to aid doctors in diagnosis [86]. These points demonstrate that the attention model is robust, flexible, applicable to a variety of fields beyond auditory attention, all the while matching human acoustic perception.

Temporal build-up was highlighted throughout this work in different ways. We first saw the effect of time in the behavioral experiment in Chapter 3, where salient events went unnoticed much more frequently if they occurred near the start of a scene. This finding indicates that the auditory system builds regularities over time, and a tolerance for deviance is much greater before regularity representations have had time to stabilize. Complementing this build-up of global attention, we also found in Chapter 5 that following attentional reset, even directed attention needs time to reach stability. Similar to the first experiment, salient events went unnoticed much more frequently if they occurred closely after an attention reorientation caused by feature shifts. Interestingly, however, this build-up only happened for shifts near the start of trials. When subjects sufficiently adapted to the scene, and thus adapted to feature changes, a build-up no longer occurred. Finally, we demonstrated that this build-up depends on the type of attention being deployed (fastest build-up for object-based attention, slowest for global attention), and is also modulated by saliency (most

prominent build-up is observed for low-saliency targets).

Tying back to the essential difference of audition compared to vision, a key take-away message from these findings is: The fact that sound evolves over time is reflected throughout a variety of attentional measures as a significant determinant of perception. This is especially the case for auditory saliency, further corroborating that adaptation of visual attention models to audition is not likely to accurately represent auditory perception. Just as the hypothetical example of listening to Haydn's suprise concerto we described in Figure 2.2, our experiments reveal that changing the placement of a sound by even 500 ms in an auditory scene can have measurably different effects on how that sound is perceived (see Figure 5.4).

Complementary to the behavioral and EEG measures of saliency-based bottom-up auditory attention, we examined how saliency affects top-down attention. Results obtained suggest that at least three different types of attention exist in the auditory pathway, namely global (free-listening, which was also used in Chapter 3), feature-based, and object-based. All of these attention types behave differently as salience of sounds is increased. Specifically, as the focus of our attention narrows, we are much more sensitive to small deviations in sound saliency. Our findings point to object-based attention resulting in a greater enhancement in perception compared to feature-based attention. As there are very few studies in both visual and auditory literatures that contrast the two types of selective attention for the same stimuli, whether this effect is restricted to the auditory modality, and whether it depends on

task demands or is a stable effect, remains to be seen.

Overall, this dissertation advances our understanding of auditory attention, including markers of bottom-up attention, methodologies of probing saliency, computationally modeling it, types of top-down attention, and how bottom-up and top-down attention interact. We have discussed why it is crucial to consider auditory attention as different from visual attention, and the wide applicability of attention models in a variety of fields. Ultimately, this work lays a foundation for understanding how auditory attention to natural scenes is processed in the brain, and the steps we can take to further uncover the properties of this information filtering mechanism.

6.1 Future work

The field of modeling auditory attention remains in its infancy. As a result, there is a large variety of avenues open for exploration, both for bottom-up and top-down attentional control. Perhaps the most urgent work the field currently needs is a set of ground-truth auditory saliency data. Some attempts are being made to collect human annotations on continuous natural scenes to determine saliency level among time [45, 75]. The central problem remains the top-down confound in having subjects actively listen to a scene. Even if distraction tasks are used, the saliency level recorded is likely to vary based on task demands, thus not representing stable ground-truth. Some of the EEG markers we extracted in Chapter 4 could be used

to derive an estimate of perceived salience over short time windows. While the time resolution of this approach is not very high, we were able to see coherence effects for windows as short as 100 ms. Without directed attention, neural responses are likely to entrain to an average of slow modulations in the scene, thus it is possible that similar metrics can be observed from more complex natural scenes with unattending subjects. Although such data would not be as easy to collect and interpret as eye-tracking data, once processed, it would make building and comparing auditory saliency models significantly easier.

The computational model can be extended in a variety of ways. The greatest advantage of the model is its flexibility; under the same framework, many of the individual components can be replaced as necessary. For example, different sound features can be used if there is a priori knowledge of informative features for a dataset, and there are endless optimization possibilities to train the weights between features based on desired tuning of the model. Although the presented model reflects global attention, it is easy to conceptualize how we could extend it to factor for feature-based attention: Boosting the weights of the desired features, or adding optimization constraints to give higher weight to selected features would most likely help to represent feature-based attention. Incorporating object-based attention into the model is less straightforward, as a separate mechanism is necessary to segregate feature streams into objects. However, if we assume the existence of a feature-integration module, it can be directly incorporated into the same predictive coding framework to find which

objects in the scene are most salient.

Despite evidence that auditory attention can be deployed in an object-based manner as demonstrated by enhanced neural representations to attended speech, and feature-based enhancement throughout the auditory pathway revealed by imaging studies, it is still unclear to what extent these forms of top-down attention differ. Our results provided behavioral evidence in support of the theory that the two forms of attention engage different mechanisms. Future experiments are necessary to determine whether the difference observed in our work is caused by suppression of unattended features as is hypothesized to be the case for visual feature-based attention. So far, the main focus of experimental studies have been to confirm attentional modulation in the auditory system. Imaging experiments with designs targeted to test suppression of unattended features are likely to clarify this possibility and enhance our understanding of auditory attention.

Bibliography

[1] E. C. Cherry, *On human communication.* Cambridge, MA: MIT Press, 1957.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] A. S. Bregman, *Auditory scene analysis: the perceptual organization of sound.* Cambridge, Mass.: MIT Press, 1990.

[4] C. Alain and L. J. Bernstein, "From sounds to meaning: the role of attention during auditory scene analysis," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 16, no. 5, pp. 485–489, 2008.

[5] J. Driver, "A selective review of selective attention research from the past century," *British Journal of Psychology*, vol. 92, no. 1, pp. 53–78, 2001.

[6] E. Awh, A. V. Belopolsky, and J. Theeuwes, "Top-down versus bottom-up attentional control: a failed theoretical dichotomy," *Trends in cognitive sciences*, vol. 16, no. 8, pp. 437–443, 2012.

BIBLIOGRAPHY

[7] L. Whiteley and M. Sahani, "Attention in a bayesian framework," *Frontiers in Human Neuroscience*, vol. 6, no. 100, 2012.

[8] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention–focusing the searchlight on sound," *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[9] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[10] I. P. Jaaskelainen and J. Ahveninen, "Auditory-cortex short-term plasticity induced by selective attention," *Neural plasticity*, vol. 2014, p. 216731, 2014.

[11] S. Shamma and J. Fritz, "Adaptive auditory computations," *Current opinion in neurobiology*, vol. 25, pp. 164–168, Apr 2014.

[12] N. M. Weinberger, *Receptive Field Plasticity and Memory in the Auditory Cortex: Coding the Learned Importance of Events*, ser. Model Systems and the neurobiology of associative learning. Lawrence Erlbaum Associates, 2001.

[13] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in a1?" *Hearing Research*, vol. 229, no. 1-2, pp. 186–203, 2007.

[14] K. T. Hill and L. M. Miller, "Auditory attentional control and selection during

cocktail party listening," *Cerebral cortex (New York, N.Y.: 1991)*, vol. 20, no. 3, pp. 583–590, Mar 2010.

[15] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[16] T. S. Lee and D. Mumford, "Hierarchical bayesian inference in the visual cortex," *The journal of the Optical Society of America*, vol. 20, no. 7, pp. 1434–1448, 2003.

[17] M. Corbetta, G. Patel, and G. L. Shulman, "The reorienting system of the human brain: From environment to theory of mind," *Neuron*, vol. 58, no. 3, pp. 306–324, 2008.

[18] E. I. Knudsen, "Fundamental components of attention," *Annu Rev Neurosci*, vol. 30, pp. 57–78, 2007.

[19] A. Yaron, I. Hershenhoren, and I. Nelken, "Sensitivity to complex statistical regularities in rat auditory cortex," *Neuron*, vol. 76, no. 3, pp. 603–615, Nov 8 2012.

[20] S. Norman-Haignere, N. G. Kanwisher, and J. H. McDermott, "Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition," *Neuron*, vol. 88, no. 6, pp. 1281–1296, Dec 16 2015.

BIBLIOGRAPHY

[21] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in cognitive sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[22] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends in neurosciences*, vol. 34, no. 3, pp. 114–123, Mar 2011.

[23] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[24] M. Carrasco, "Visual attention: The past 25 years," *Vision research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[25] Z. Li, "A saliency map in primary visual cortex," *Trends in cognitive sciences*, vol. 6, no. 1, pp. 9–16, Jan 1 2002.

[26] J. Bullier, "Integrated model of visual processing," *Brain research reviews*, vol. 36, no. 2-3, pp. 96–107, Oct 2001.

[27] R. Veale, Z. M. Hafed, and M. Yoshida, "How is visual salience computed in the brain? insights from behaviour, neurobiology and modelling," *Philosophical transactions of the Royal Society of London.Series B, Biological sciences*, vol. 372, no. 1714, p. 10.1098/rstb.2016.0113. Epub 2017 Jan 2, February 19 2017.

[28] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 185–207, 2013.

[29] J. Zheng, J. Tian, K. Deng, X. Dai, X. Zhang, and M. Xu, "Salient feature region: A new method for retinal image registration," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 221–232, 2011.

[30] T. V. Nguyen, Z. Song, and S. Yan, "Stap: Spatial-temporal attention-aware pooling for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 1, pp. 77–86, 2015.

[31] Y. Yu, G. K. Mann, and R. G. Gosine, "Target tracking for moving robots using object-based visual attention," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*. IEEE, 2010, pp. 2902–2907.

[32] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *INTERSPEECH-2007*, 2007, pp. 1941–1944.

[33] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," in *15th European Signal Processing Conference (EUSIPCO 2007)*, 2007.

[34] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis

of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[35] E. M. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, 2012.

[36] T. Tsuchida and G. Cottrell, "Auditory saliency using natural statistics," 2012.

[37] L. Zhang, M. H. Tong, and T. K. Marks, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, p. 32, 2008.

[38] E. M. Kaya and M. Elhilali, "Investigating bottom-up auditory attention," *Frontiers in Human neuroscience*, vol. 8, no. 327, p. doi: 10.3389/fnhum.2014.00327, 2014.

[39] I. Winkler, "Interpreting the mismatch negativity." *Journal of Psychophysiology*, vol. 21, no. 3-4, p. 147, 2007.

[40] K. J. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[41] R. Southwell, A. Baumann, C. Gal, N. Barascud, K. Friston, and M. Chait, "Is predictability salient? a study of attentional capture by auditory patterns," *Philosophical transactions of the Royal Society of London.Series B, Biological*

sciences, vol. 372, no. 1714, p. 10.1098/rstb.2016.0105. Epub 2017 Jan 2, Feb 19 2017.

[42] J. Wang, K. Zhang, K. Madani, and C. Sabourin, "Salient environmental sound detection framework for machine awareness," *Neurocomputing*, vol. 152, pp. 444–454, 2015.

[43] P. Mermelstein, *Distance measures for speech recognition, psychological and instrumental*, ser. Pattern Recognition and Artificial Intelligence. Academic, New York, 1976, pp. 374–388.

[44] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[45] K. Kim, K.-H. Lin, D. B. Walther, M. A. Hasegawa-Johnson, and T. S. Huang, "Automatic detection of auditory salience with optimized linear filters derived from human annotation," *Pattern Recognition Letters*, vol. 38, no. 0, pp. 78–85, 2014.

[46] F. Tordini, A. S. Bregman, J. R. Cooperstock, A. Ankolekar, and T. Sandholm, "Toward an improved model of auditory saliency," in *Proceedings of the 19th International Conference on Auditory Display (ICAD2013)*, 2013 2013.

[47] F. Tordini, A. S. Bregman, and J. R. Cooperstock, "The loud bird doesn't (al-

ways) get the worm: Why computational salience also needs brightness and tempo," in *Proceedings of the 21st International Conference on Auditory Display (ICAD 2015)*, 2015 2015.

[48] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *Journal of Vision*, vol. 11, no. 5, May 27 2011.

[49] J. Stoll, M. Thrun, A. Nuthmann, and W. Einhauser, "Overt attention in natural scenes: objects dominate features," *Vision research*, vol. 107, pp. 36–48, Feb 2015.

[50] N. M. Weinberger, *Reconceptualizing the primary auditory cortex: Learning, memory and specific plasticity*, ser. The Auditory Cortex. New York: Springer, 2011, ch. Chapter 22, pp. 465–491.

[51] D. H. Hubel, C. O. Henson, A. Rupert, and R. Galambos, "Attention units in the auditory cortex," *Science*, vol. 129, no. 3358, pp. 1279–1280, 1959.

[52] M. Elhilali, S. A. Shamma, J. Z. Simon, and J. B. Fritz, *A Linear Systems View to the Concept of STRF*, ser. Handbook of Modern Techniques in Auditory Cortex. Nova Science Pub Inc, 2013, pp. 33–60.

[53] A. M. H. J. Aertsen and P. I. M. Johannesma, "The spectro-temporal receptive field," *Biological Cybernetics*, vol. 42, pp. 133–143, 1981.

BIBLIOGRAPHY

[54] S. Shamma, "Characterizing auditory receptive fields," *Neuron*, vol. 58, no. 6, pp. 829–831, 2008.

[55] P. Yin, J. B. Fritz, and S. A. Shamma, "Rapid spectrotemporal plasticity in primary auditory cortex during behavior," *The Journal of neuroscience*, vol. 34, no. 12, pp. 4396–4408, Mar 19 2014.

[56] S. V. David, J. B. Fritz, and S. A. Shamma, "Task reward structure shapes rapid receptive field plasticity in auditory cortex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 6, pp. 2144–2149, Feb 7 2012.

[57] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Adaptive changes in cortical receptive fields induced by attention to complex sounds," *Journal of Neurophysiology*, vol. 98, no. 4, pp. 2337–2346, 2007.

[58] P. Zurita, A. E. Villa, Y. de Ribaupierre, F. de Ribaupierre, and E. M. Rouiller, "Changes of single unit activity in the cat's auditory thalamus and cortex associated to different anesthetic conditions," *Neuroscience Research*, vol. 19, no. 3, pp. 303–316, 1994.

[59] S. Atiani, S. V. David, D. Elgueda, M. Locastro, S. Radtke-Schuller, S. A. Shamma, and J. B. Fritz, "Emergent selectivity for task-relevant stimuli in higher-order auditory cortex," *Neuron*, vol. 82, no. 2, pp. 486–499, Apr 16 2014.

[60] C. I. Petkov, X. Kang, K. Alho, O. Bertrand, E. W. Yund, and D. L. Woods, "Attentional modulation of human auditory cortex," *Nature neuroscience*, vol. 7, no. 6, pp. 658–663, Jun 2004.

[61] C. Alain and D. L. Woods, "Attention modulates auditory pattern memory as indexed by event-related brain potentials," *Psychophysiology*, vol. 34, no. 5, pp. 534–546, Sep 1997.

[62] A. K. Lee, S. Rajaram, J. Xia, H. Bharadwaj, E. Larson, M. S. Hämäläinen, and B. G. Shinn-Cunningham, "Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch," *Frontiers in neuroscience*, vol. 6, 2012.

[63] R. J. Zatorre, T. A. Mondor, and A. C. Evans, "Auditory attention to space and frequency activates similar cerebral systems," *NeuroImage*, vol. 10, no. 5, pp. 544–554, Nov 1999.

[64] J. Ahveninen, I. P. Jaaskelainen, T. Raij, G. Bonmassar, S. Devore, M. Hamalainen, S. Levanen, F. H. Lin, M. Sams, B. G. Shinn-Cunningham, T. Witzel, and J. W. Belliveau, "Task-modulated "what" and "where" pathways in human auditory cortex," *Proceedings of the National Academy of Sciences*, vol. 103, no. 39, pp. 14 608–14 613, 2006.

[65] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proceedings of the National Academy of*

Sciences of the United States of America, vol. 109, no. 29, pp. 11 854–11 859, 2012.

[66] N. Mesgarani, J. Fritz, and S. Shamma, "A computational model of rapid task-related plasticity of auditory cortical receptive fields," *Journal of computational neuroscience*, vol. 28, no. 1, pp. 19–27, Feb 2010.

[67] M. Carlin and M. Elhilali, "A framework for speech activity detection using adaptive auditory receptive fields," *IEEE Transactions on Audio Speech and Language Processing*, vol. 23, no. 12, pp. 2422–2433, 2015.

[68] O. Kalinli and S. Narayanan, "Combining task-dependent information with auditory attention cues for prominence detection in speech," in *9th Annual Conference of the International-Speech-Communication-Association*, 2008, pp. 1064–1067.

[69] K. Patil and M. Elhilali, "Task-driven attentional mechanisms for auditory scene recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 828–832.

[70] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial eeg," *Cerebral Cortex*, January 15 2014.

BIBLIOGRAPHY

[71] B. Mirkovic, S. Debener, M. Jaeger, and M. D. Vos, "Decoding the attended speech stream with multi-channel eeg: implications for online, daily-life applications," *Journal of neural engineering*, vol. 12, no. 4, p. 046007, Aug 2015.

[72] S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi, "Robust decoding of selective auditory attention from meg in a competing-speaker environment via state-space modeling," *NeuroImage*, vol. 124, Part A, pp. 906–917, 2016.

[73] M. Yang, S. A. Sheth, C. A. Schevon, G. M. M. II, and N. Mesgarani, "Speech reconstruction from human auditory cortex with deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, September 2015 2015.

[74] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, nov 1998.

[75] N. Huang and M. Elhilali, "Around the world in 20 scenes: An exploration of auditory saliency through a selection of complex natural scenes," 2015.

[76] R. M. Cichy and S. Teng, "Resolving the neural dynamics of visual and auditory scene processing in the human brain: a methodological approach," *Phil. Trans. R. Soc. B*, vol. 372, no. 1714, p. 20160108, 2017.

[77] O. Kalinli, S. Sundaram, and S. Narayanan, "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing," in *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on.* IEEE, 2009, pp. 1–6.

[78] M. Slaney, T. Agus, S.-C. Liu, M. Kaya, and M. Elhilali, "A model of attention-driven scene analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 145–148.

[79] M. A. Carlin and M. Elhilali, "Modeling attention-driven plasticity in auditory cortical receptive fields," *Frontiers in computational neuroscience*, vol. 9,, p. doi: 10.3389/fncom.2015.00106, Aug 19 2015.

[80] A. Bellur and M. Elhilali, "Feedback driven sensory mapping adaptation for robust speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[81] D. Oldoni, B. D. Coensel, M. Rademaker, B. D. Baets, and D. Botteldooren, "Context-dependent environmental sound monitoring using som coupled with legion," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–8.

[82] R. Hu, B. Hang, Y. Ma, and S. Dong, "A bottom-up audio attention model for surveillance," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, 2010, pp. 564–567.

[83] T. Liu and I. Mance, "Constant spread of feature-based attention across the visual field," *Vision research*, vol. 51, no. 1, pp. 26–33, January 01 2011.

[84] S. Kakouros, O. Räsänen, and U. K. Laine, "Attention based temporal filtering of sensory signals for data redundancy reduction," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3188–3192.

[85] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[86] E. M. Kaya and M. Elhilali, "Abnormality detection in noisy biosignals," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 3949–3952.

[87] J. Nakajima, A. Sugimoto, and K. Kawamoto, *Incorporating Audio Signals into Constructing a Visual Saliency Map*, ser. Image and Video Technology. Springer Berlin Heidelberg, 2014, pp. 468–480.

[88] N. O. Sidaty, M.-C. Larabi, and A. Saadane, "An audiovisual saliency model for conferencing and conversation videos," in *IST International Symposium on Electronic Imaging*, vol. 2016, 2016.

[89] S. Ramenahalli, D. R. Mendat, S. Dura-Bernal, E. Culurciello, E. Niebur, and

BIBLIOGRAPHY

A. Andreou, "Audio-visual saliency map: Overview, basic models and hardware implementation," in *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, 2013, pp. 1–6.

[90] R. J. Duro, F. Bellas, and J. A. B. Permuy, *Brain-Like Robotics*, ser. Springer Handbook of Bio-/Neuroinformatics. Berlin, Heidelberg: Springer, 2014, vol. 2016, ch. 5/15/2016, pp. 1019–1056.

[91] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, 2008, pp. 962–967.

[92] J. L. Crespo, A. Faiña, and R. J. Duro, "An adaptive detection/attention mechanism for real time robot operation," *Neurocomputing*, vol. 72, no. 4-6, p. 850 ¡last_page¿ 860, 2009.

[93] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[94] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[95] C. M. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows

what is interesting: Salient locations which should be fixated," *Journal of Vision*, vol. 9, no. 11, pp. 1–22, 2009.

[96] J. M. Wolfe, M. L. H. Vo, K. K. Evans, and M. R. Greene, "Visual search in scenes involves selective and non-selective pathways," *Trends in Cognitive Sciences*, vol. 15, no. 2, pp. 77–84, 2011.

[97] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[98] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, p. 15, 2009.

[99] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[100] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 4, pp. 996–1010, 2012.

[101] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, p. 5, 2009.

[102] A. Ihlefeld and B. Shinn-Cunningham, "Disentangling the effects of spatial cues

on selection and formation of auditory objects a," *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2224–2235, 2008.

[103] R. Naatanen, A. W. Gaillard, and S. Mantysalo, "Early selective-attention effect on evoked potential reinterpreted," *Acta Psychologica*, vol. 42, no. 4, pp. 313–329, 1978.

[104] P. May and H. Tiitinen, "Mismatch negativity (mmn), the deviance-elicited auditory deflection, explained," *Psychophysiology*, vol. 47, no. 1, pp. 66–122, 2010.

[105] D. C. Knill and A. Pouget, "The bayesian brain: the role of uncertainty in neural coding and computation," *Trends in neurosciences*, vol. 27, no. 12, pp. 712–719, 2004.

[106] K. Friston, "A theory of cortical responses," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1456, pp. 815–836, April 29 2005.

[107] M. I. Garrido, J. M. Kilner, K. E. Stephan, and K. J. Friston, "The mismatch negativity: A review of underlying mechanisms," *Clinical Neurophysiology*, vol. 120, no. 3, p. 453, 2009.

[108] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the

allocation of overt visual attention," *Vision research*, vol. 42, no. 1, pp. 107–123, 2002.

[109] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? a study of human explicit saliency judgment," *Vision research*, vol. 91, pp. 62–77, 2013.

[110] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE transactions on information theory*, vol. 38, no. 2, pp. 824–839, 1992.

[111] R. D. Melara and L. E. Marks, "Interaction among auditory dimensions: timbre, pitch, and loudness," *Attention, Perception, & Psychophysics*, vol. 48, no. 2, pp. 169–178, 1990.

[112] E. J. Allen and A. J. Oxenham, "Interactions of pitch and timbre: How changes in one dimension affect discrimination of the other," in *Abstracts of the Thirty-sixth ARO Mid-Winter meeting, Volume 36, Mt. Royal, NJ: Association of Research Otolaryngologists*, vol. 36, 2013 2013.

[113] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[114] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database:

Music genre database and musical instrument sound database," *Proceedings of International Symposium on Music Information Retrieval*, pp. 229–230, 2003.

[115] S. A. Shamma and D. J. Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2631–2644, 2000.

[116] K. M. Walker, J. K. Bizley, A. J. King, and J. W. Schnupp, "Multiplexed and robust representations of sound features in auditory cortex," *Journal of Neuroscience*, vol. 31, no. 41, pp. 14 565–14 576, 2011.

[117] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: the biological bases of musical timbre perception," *PLoS Comput Biol*, vol. 8, no. 11, p. e1002759, 2012.

[118] Z. Chen, "Bayesian filtering: From kalman filters to particle filters, and beyond," *Statistics*, vol. 182, no. 1, pp. 1–69, 2003.

[119] E. Arnaud, E. Memin, and B. Cernuschi-Frias, "Conditional filters for image sequence-based tracking - application to point tracking," *IEEE Transactions on Image Processing*, vol. 14, no. 1, pp. 63–79, 2005.

[120] S. McAdams, S. Winsberg, S. Donnadieu, G. D. Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: common dimensions, specifici-

ties, and latent subject classes," *Psychological Research*, vol. 58, no. 3, pp. 177–192, 1995.

[121] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. Emerald Group Publishing Ltd., 2003.

[122] I. Winkler, S. L. Denham, and I. Nelken, "Modeling the auditory scene: predictive regularity representations and perceptual objects," *Trends in cognitive sciences*, vol. 13, no. 12, p. 40, 2009.

[123] A. Bendixen, S. L. Denham, K. Gyimesi, and I. Winkler, "Regular patterns stabilize auditory streams," *Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3658–3666, 2010.

[124] L.-V. Andreou, M. Kashino, and M. Chait, "The role of temporal regularity in auditory segregation," *Hearing research*, vol. 280, no. 1–2, pp. 228–235, 2011.

[125] T. Rahne and E. Sussman, "Neural representations of auditory input accommodate to the context in a dynamically changing acoustic environment," *The European journal of neuroscience*, vol. 29, no. 1, pp. 205–211, Jan 2009.

[126] A. S. Bregman, "Auditory streaming is cumulative," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 4, no. 3, pp. 380–387, 1978.

BIBLIOGRAPHY

[127] S. Anstis and S. Saida, "Adaptation to auditory streaming of frequency-modulated tones," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 11, no. 3, pp. 257–271, 1985.

[128] N. R. Haywood and B. Roberts, "Build-up of the tendency to segregate auditory streams: Resetting effects evoked by a single deviant tone," *Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3019–3031, 2010.

[129] C. Micheyl, B. Tian, R. P. Carlyon, and J. P. Rauschecker, "Perceptual organization of tone sequences in the auditory cortex of awake macaques," *Neuron*, vol. 48, no. 1, pp. 139–148, 2005.

[130] D. Pressnitzer, M. Sayles, C. Micheyl, and I. M. Winter, "Perceptual organization of sound begins in the auditory periphery," *Current Biology*, vol. 18, no. 15, pp. 1124–1128, 2008.

[131] H. M. Kondo, N. Kitagawa, M. S. Kitamura, A. Koizumi, M. Nomura, and M. Kashino, "Separability and commonality of auditory and visual bistable perception," *Cerebral Cortex*, vol. 22, no. 8, pp. 1915–1922, August 01 2012.

[132] E. S. Sussman, "A new view on the mmn and attention debate," *Journal of Psychophysiology*, vol. 21, no. 3, pp. 164–175, 2007.

[133] T. W. Picton, C. Alain, L. Otten, W. Ritter, and A. Achim, "Mismatch neg-

ativity: Different water in the same river." *Audiology and Neurotology*, vol. 5, pp. 111–139, 2000.

[134] M. Garagnani and F. Pulvermuller, "From sounds to words: a neurocomputational model of adaptation, inhibition and memory processes in auditory change detection," *NeuroImage*, vol. 54, no. 1, pp. 170–181, 2011.

[135] A. Bendixen, I. SanMiguel, and E. Schroger, "Early electrophysiological indicators for predictive processing in audition: A review," *Psychophysiology*, vol. 83, no. 2, pp. 120–131, 2012.

[136] F. Lieder, J. Daunizeau, M. I. Garrido, K. J. Friston, and K. E. Stephan, "Modelling trial-by-trial changes in the mismatch negativity," *PLoS computational biology*, vol. 9, no. 2, p. e1002911, 2013.

[137] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston, "Canonical microcircuits for predictive coding," *Neuron*, vol. 76, no. 4, pp. 695–711, 2012.

[138] R. Linsker, "Neural network learning of optimal kalman prediction and control," *Neural networks*, vol. 21, no. 9, pp. 1328–1343, 2008.

[139] G. Szirtes, B. Poczos, and A. Lorincz, "Neural kalman filter," *Neurocomputing*, vol. 65-66, pp. 349–355, 2005.

[140] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004.

[141] D. T. Mirikitani and N. Nikolaev, "Recursive bayesian recurrent neural networks for time-series modeling," *IEEE Transactions on Neural Networks*, vol. 21, no. 2, pp. 262–274, 2010.

[142] M. Korenberg and I. Hunter, "The identification of nonlinear biological systems: Volterra kernel approaches," *Annals of Biomedical Engineering*, vol. 24, no. 4, pp. 250–268, 1996.

[143] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2006.

[144] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: a bayesian inference theory of attention," *Vision research*, vol. 50, pp. 2233–2247, 2010.

[145] M. W. Spratling, "Predictive coding accounts for v1 response properties recorded using reverse correlation," *Biological Cybernetics*, vol. 106, no. 1, pp. 37–49, 2012.

[146] M. Kimura, "Visual mismatch negativity and unintentional temporal-context-based prediction in vision," *International Journal of Psychophysiology*, vol. 83, no. 2, pp. 144–155, 2012.

[147] K. Akatsuka, T. Wasaka, H. Nakata, T. Kida, and R. Kakigi, "The effect of

stimulus probability on the somatosensory mismatch field," *Experimental Brain Research*, vol. 181, no. 4, pp. 607–614, 2007.

[148] M. Sabri, A. J. Radnovich, T. Q. Li, and D. A. Kareken, "Neural correlates of olfactory change detection," *NeuroImage*, vol. 25, no. 3, pp. 969–974, 2005.

[149] M. G. Woldorff, C. C. Gallen, S. A. Hampson, S. A. Hillyard, C. Pantev, D. Sobel, and F. E. Bloom, "Modulation of early sensory processing in human auditory cortex during auditory selective attention," *Proceedings of the National Academy of Sciences*, vol. 90, no. 18, pp. 8722–8726, 1993.

[150] N. Huang and M. Elhilali, "Auditory salience using natural soundscapes," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2163–2176, 2017.

[151] E. M. Kaya and M. Elhilali, "Modelling auditory attention," *Philosophical transactions of the Royal Society of London.Series B, Biological sciences*, vol. 372, no. 1714, p. 10.1098/rstb.2016.0101, Feb 19 2017.

[152] G. H. Bishop, "Cyclic changes in excitability of the optic pathway of the rabbit," *Am J Physiol*, vol. 103, no. 1, p. 213, 1932.

[153] L. M. Ward, "Synchronous neural oscillations and cognitive processes," *Trends in cognitive sciences*, vol. 7, no. 12, pp. 553–559, December 01 2003.

[154] G. Thut and C. Miniussi, "New insights into rhythmic brain activity from tms-eeg studies," *Trends in cognitive sciences*, vol. 13, no. 4, pp. 182–189, Apr 2009.

[155] S. Makeig, S. Debener, J. Onton, and A. Delorme, "Mining event-related brain dynamics," *Trends in cognitive sciences*, vol. 8, no. 5, pp. 204–210, May 01 2004.

[156] S. Luck, *An Introduction to the Event-Related Potential Technique*. MIT Press, 2005.

[157] T. W. Picton, S. A. Hillyard, H. I. Krausz, and R. Galambos, "Human auditory evoked potentials. i. evaluation of components," *Electroencephalography and clinical neurophysiology*, vol. 36, no. 2, pp. 179–190, February 01 1974.

[158] R. Naatanen, P. Paavilainen, T. Rinne, and K. Alho, "The mismatch negativity (mmn) in basic research of central auditory processing: a review," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 118, no. 12, pp. 2544–2590, December 01 2007.

[159] G. Stefanics, P. Astikainen, and I. Czigler, "Visual mismatch negativity (vmmn): a prediction error signal in the visual modality," *Frontiers in human neuroscience*, vol. 8, p. 1074, January 22 2015.

[160] G. Thut, P. G. Schyns, and J. Gross, "Entrainment of perceptually relevant

brain oscillations by non-invasive rhythmic stimulation of the human brain," *Frontiers in psychology*, vol. 2, p. 170, July 20 2011.

[161] J. Besle, C. A. Schevon, A. D. Mehta, P. Lakatos, R. R. Goodman, G. M. McKhann, R. G. Emerson, and C. E. Schroeder, "Tuning of the human neocortex to the temporal dynamics of attended events," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 31, no. 9, pp. 3176–3185, March 02 2011.

[162] T. Neuling, S. Rach, S. Wagner, C. H. Wolters, and C. S. Herrmann, "Good vibrations: oscillatory phase shapes perception," *NeuroImage*, vol. 63, no. 2, pp. 771–778, November 01 2012.

[163] P. Lakatos, G. Karmos, A. D. Mehta, I. Ulbert, and C. E. Schroeder, "Entrainment of neuronal oscillations as a mechanism of attentional selection," *Science (New York, N.Y.)*, vol. 320, no. 5872, pp. 110–113, Apr 4 2008.

[164] M. Elhilali, J. Xiang, S. Shamma, and J. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLOS Biology*, vol. 7, no. 6, p. e1000129, 2009.

[165] P. Lakatos, G. Musacchia, M. N. O'Connel, A. Y. Falchier, D. C. Javitt, and C. E. Schroeder, "The spectrotemporal filter mechanism of auditory selective attention," *Neuron*, vol. 77, no. 4, pp. 750–761, February 20 2013.

[166] L. Shuai and M. Elhilali, "Task-dependent neural representations of salient events in dynamic auditory scenes," *Frontiers in neuroscience*, vol. 8, no. 203, p. 10.3389/fnins.2014.00203, 2014.

[167] C. E. Schroeder and P. Lakatos, "Low-frequency neuronal oscillations as instruments of sensory selection," *Trends in neurosciences*, vol. 32, no. 1, pp. 9–18, 2009.

[168] G. Stefanics, B. Hangya, I. Hernadi, I. Winkler, P. Lakatos, and I. Ulbert, "Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, no. 41, pp. 13 578–13 585, Oct 13 2010.

[169] H. Luo and D. Poeppel, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron*, vol. 54, no. 6, pp. 1001–1010, June 21 2007.

[170] S. J. Aiken and T. W. Picton, "Human cortical responses to the speech envelope," *Ear and hearing*, vol. 29, no. 2, pp. 139–157, April 01 2008.

[171] M. F. Howard and D. Poeppel, "Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension," *Journal of neurophysiology*, vol. 104, no. 5, pp. 2500–2511, November 01 2010.

[172] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M.

McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, and C. E. Schroeder, "Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"," *Neuron*, vol. 77, no. 5, pp. 980–991, Mar 6 2013.

[173] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations," *Frontiers in human neuroscience*, vol. 8, p. 311, May 28 2014.

[174] B. S. Ng, T. Schroeder, and C. Kayser, "A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 32, no. 35, pp. 12 268–12 276, August 29 2012.

[175] M. Steinschneider, K. V. Nourski, and Y. I. Fishman, "Representation of speech in human auditory cortex: is it special?" *Hearing research*, vol. 305, pp. 57–73, November 01 2013.

[176] B. S. Ng, N. K. Logothetis, and C. Kayser, "Eeg phase patterns reflect the selectivity of neural firing," *Cerebral cortex (New York, N.Y.: 1991)*, vol. 23, no. 2, pp. 389–398, February 01 2013.

[177] R. Oostenveld, P. Fries, E. Maris, and J. M. Schoffelen, "Fieldtrip: Open source software for advanced analysis of meg, eeg, and invasive electrophysiological data," *Computational intelligence and neuroscience*, vol. 2011, p. 156869, 2011.

[178] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[179] S. Krstulovic and R. Gribonval, "Mptk: Matching pursuit made tractable," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 3. IEEE, 2006, p. III.

[180] S. V. David, N. Mesgarani, and S. A. Shamma, "Estimating sparse spectro-temporal receptive fields with natural stimuli," *Network: Computation in Neural Systems*, vol. 18, no. 3, pp. 191–212, 2007.

[181] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of neurophysiology*, vol. 107, no. 1, pp. 78–89, January 01 2012.

[182] J. R. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a "cocktail party"," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, no. 2, pp. 620–628, January 13 2010.

[183] M. J. Henry and J. Obleser, "Frequency modulation entrains slow neural oscillations and optimizes human listening behavior," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 49, pp. 20 095–20 100, December 04 2012.

[184] X. Wang, T. Lu, D. Bendor, and E. Bartlett, "Neural coding of temporal information in auditory thalamus and cortex," *Neuroscience*, vol. 154, no. 1, pp. 294–303, 2008.

[185] C. Kayser, M. A. Montemurro, N. K. Logothetis, and S. Panzeri, "Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns," *Neuron*, vol. 61, no. 4, pp. 597–608, 2009.

[186] C. Chandrasekaran, H. K. Turesson, C. H. Brown, and A. A. Ghazanfar, "The influence of natural scene dynamics on auditory cortical activity," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, no. 42, pp. 13 919–13 931, October 20 2010.

[187] A. Muller-Gass, M. Macdonald, E. Schröger, L. Sculthorpe, and K. Campbell, "Evidence for the auditory p3a reflecting an automatic process: elicitation during highly-focused continuous visual attention," *Brain research*, vol. 1170, pp. 71–78, 2007.

[188] C. Escera and M. Corral, "Role of mismatch negativity and novelty-p3 in involuntary auditory attention," *Journal of Psychophysiology*, vol. 21, no. 3-4, pp. 251–264, 2007.

[189] C. Escera, K. Alho, E. Schröger, and I. W. Winkler, "Involuntary attention and distractibility as evaluated with event-related brain potentials," *Audiology and Neurotology*, vol. 5, no. 3-4, pp. 151–166, 2000.

BIBLIOGRAPHY

[190] J. Polich, "Updating p300: an integrative theory of p3a and p3b," *Clinical neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.

[191] O. David, L. Harrison, and K. J. Friston, "Modelling event-related responses in the brain," *NeuroImage*, vol. 25, no. 3, pp. 756–770, 2005.

[192] P. Sauseng, W. Klimesch, W. R. Gruber, S. Hanslmayr, R. Freunberger, and M. Doppelmayr, "Are event-related potential components generated by phase resetting of brain oscillations? a critical discussion," *Neuroscience*, vol. 146, no. 4, pp. 1435–1444, June 08 2007.

[193] W. Klimesch, M. Schabus, M. Doppelmayr, W. Gruber, and P. Sauseng, "Evoked oscillations and early components of event-related potentials: an analysis," *International Journal of Bifurcation and Chaos*, vol. 14, no. 02, pp. 705–718, 2004.

[194] L. Fuentemilla, J. Marco-Pallares, T. F. Munte, and C. Grau, "Theta eeg oscillatory activity and auditory change detection," *Brain research*, vol. 1220, pp. 93–101, July 18 2008.

[195] F. J. Hsiao, Z. A. Wu, L. T. Ho, and Y. Y. Lin, "Theta oscillation during auditory change detection: An meg study," *Biological psychology*, vol. 81, no. 1, pp. 58–66, April 01 2009.

[196] G. Stothart and N. Kazanina, "Oscillatory characteristics of the visual mis-

match negativity: what evoked potentials aren't telling us," *Frontiers in human neuroscience*, vol. 7, p. 426, August 01 2013.

[197] M. Xu, Y. Jia, H. Qi, Y. Hu, F. He, X. Zhao, P. Zhou, L. Zhang, B. Wan, W. Gao, and D. Ming, "Use of a steady-state baseline to address evoked vs. oscillation models of visual evoked potential origin," *NeuroImage*, vol. 134, pp. 204–212, July 01 2016.

[198] S. Treue and J. H. Maunsell, "Attentional modulation of visual motion processing in cortical areas mt and mst," *Nature*, vol. 382, no. 6591, pp. 539–541, August 08 1996.

[199] M. Saenz, G. T. Buracas, and G. M. Boynton, "Global effects of feature-based attention in human visual cortex," *Nature neuroscience*, vol. 5, no. 7, pp. 631–632, July 01 2002.

[200] K. M. O'Craven, P. E. Downing, and N. Kanwisher, "fmri evidence for objects as the units of attentional selection," *Nature*, vol. 401, no. 6753, pp. 584–587, 1999.

[201] M. A. Schoenfeld, C. Tempelmann, A. Martinez, J.-M. Hopf, C. Sattler, H.-J. Heinze, and S. A. Hillyard, "Dynamics of feature binding during object-selective attention," *Proceedings of the National Academy of Sciences*, vol. 100, no. 20, pp. 11 806–11 811, 2003.

[202] W. Sohn, T. V. Papathomas, E. Blaser, and Z. Vidnyanszky, "Object-based cross-feature attentional modulation from color to motion," *Vision research*, vol. 44, no. 12, pp. 1437–1443, June 01 2004.

[203] D. Wegener, F. Ehn, M. K. Aurich, F. O. Galashan, and A. K. Kreiter, "Feature-based attention and the suppression of non-relevant object features," *Vision research*, vol. 48, no. 27, pp. 2696–2707, December 01 2008.

[204] S. Taya, W. J. Adams, E. W. Graf, and N. Lavie, "The fate of task-irrelevant visual motion: perceptual load versus feature-based attention," *Journal of vision*, vol. 9, no. 12, p. 1210, November 18 2009.

[205] E. D. Freeman, E. Macaluso, G. Rees, and J. Driver, "fmri correlates of object-based attentional facilitation vs. suppression of irrelevant stimuli, dependent on global grouping and endogenous cueing," *Frontiers in integrative neuroscience*, vol. 8, p. 12, February 10 2014.

[206] D. J. Kravitz and M. Behrmann, "Space-, object-, and feature-based attention interact to organize visual scenes," *Attention, perception & psychophysics*, vol. 73, no. 8, pp. 2434–2447, November 01 2011.

[207] K. M. Mayer and Q. C. Vuong, "The influence of unattended features on object processing depends on task demand," *Vision research*, vol. 56, pp. 20–27, 2012.

[208] D. Wegener, F. O. Galashan, M. K. Aurich, and A. K. Kreiter, "Attentional

spreading to task-irrelevant object features: experimental support and a 3-step model of attention for object-based selection and feature-based processing modulation," *Frontiers in Human Neuroscience*, vol. 8, p. 414, 2014.

[209] S. Geigerman, P. Verhaeghen, and J. Cerella, "To bind or not to bind, that's the wrong question: Features and objects coexist in visual short-term memory," *Acta Psychologica*, vol. 167, pp. 45–51, June 01 2016.

[210] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature neuroscience*, vol. 6, no. 11, pp. 1216–1223, 2003.

[211] J. Fritz, M. Elhilali, and S. Shamma, "Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex," *Hearing research*, vol. 206, no. 1-2, pp. 159–176, 2005.

[212] S. A. Hillyard, R. F. Hink, V. L. Schwent, and T. W. Picton, "Electrical signs of selective attention in the human brain," *Science*, vol. 182, no. 108, pp. 177–180, 1973.

[213] C. L. Grady, J. W. V. Meter, J. M. Maisog, P. Pietrini, J. Krasuski, and J. P. Rauschecker, "Attention-related modulation of activity in primary and secondary auditory cortex," *Neuroreport*, vol. 8, no. 11, pp. 2511–2516, July 28 1997.

[214] L. Jancke, T. W. Buchanan, K. Lutz, and N. J. Shah, "Focused and nonfocused attention in verbal and emotional dichotic listening: an fmri study," *Brain and language*, vol. 78, no. 3, pp. 349–363, September 01 2001.

[215] A. Bidet-Caulet, C. Fischer, J. Besle, P. E. Aguera, M. H. Giard, and O. Bertrand, "Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex," *Journal of Neuroscience*, vol. 27, no. 35, pp. 9252–9261, 2007.

[216] T. A. Mondor, R. J. Zatorre, and N. A. Terrio, "Constraints on the selection of auditory information." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, no. 1, p. 66, 1998.

[217] D. L. Woods, G. C. Stecker, T. Rinne, T. J. Herron, A. D. Cate, E. W. Yund, I. Liao, and X. Kang, "Functional maps of human auditory cortex: effects of acoustic features and attention," *PloS one*, vol. 4, no. 4, p. e5183, 2009.

[218] K. Krumbholz, S. B. Eickhoff, and G. R. Fink, "Feature-and object-based attentional modulation in the human auditory "where" pathway," *Journal of cognitive neuroscience*, vol. 19, no. 10, pp. 1721–1733, 2007.

[219] C. F. Altmann, M. Henning, M. K. Doring, and J. Kaiser, "Effects of feature-selective attention on auditory pattern and location processing," *NeuroImage*, vol. 41, no. 1, pp. 69–79, May 15 2008.

[220] A. Degerman, T. Rinne, A. K. Sarkka, J. Salmi, and K. Alho, "Selective attention to sound location or pitch studied with event-related brain potentials and magnetic fields," *The European journal of neuroscience*, vol. 27, no. 12, pp. 3329–3341, June 01 2008.

[221] A. E. Paltoglou, C. J. Sumner, and D. A. Hall, "Examining the role of frequency specificity in the enhancement and suppression of human cortical activity by auditory selective attention," *Hearing research*, vol. 257, no. 1-2, pp. 106–118, November 01 2009.

[222] S. D. Costa, W. van der Zwaag, L. M. Miller, S. Clarke, and M. Saenz, "Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 33, no. 5, pp. 1858–1863, January 30 2013.

[223] J. Oh, J. H. Kwon, P. S. Yang, and J. Jeong, "Auditory imagery modulates frequency-specific areas in the human auditory cortex," *Journal of cognitive neuroscience*, vol. 25, no. 2, pp. 175–187, February 01 2013.

[224] L. Riecke, J. C. Peters, G. Valente, V. G. Kemper, E. Formisano, and B. Sorger, "Frequency-selective attention in auditory scenes recruits frequency representations throughout human superior temporal cortex," *Cerebral cortex (New York, N.Y.: 1991)*, vol. 27, no. 5, pp. 3002–3014, May 01 2017.

[225] V. Best, E. J. Ozmeral, N. Kopco, and B. G. Shinn-Cunningham, "Object

continuity enhances selective auditory attention," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 35, pp. 13 174–13 178, Sep 2 2008.

[226] R. K. Maddox and B. G. Shinn-Cunningham, "Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention," *Journal of the Association for Research in Otolaryngology : JARO*, vol. 13, no. 1, pp. 119–129, February 01 2012.

[227] J. Z. Simon, "The encoding of auditory objects in auditory cortex: insights from magnetoencephalography," *International Journal of Psychophysiology*, vol. 95, no. 2, pp. 184–190, 2015.

[228] A. De Götzen, N. Bernardini, and D. Arfib, "Traditional (?) implementations of a phase vocoder: the tricks of the trade," in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, 2000.

[229] R. Cusack, J. Deeks, G. Aikman, and R. P. Carlyon, "Effects of location, frequency region, and time course of selective attention on auditory scene analysis," *Journal of experimental psychology: human perception and performance*, vol. 30, no. 4, pp. 643–656, 2004.

[230] B. Shinn-Cunningham, V. Best, and A. K. Lee, "Auditory object formation and selection," in *The Auditory System at the Cocktail Party*. Springer, 2017, pp. 7–40.

[231] B. J. Scholl, "Objects and attention: the state of the art," *Cognition*, vol. 80, no. 1-2, pp. 1–46, June 01 2001.

[232] L. Busse, K. C. Roberts, R. E. Crist, D. H. Weissman, and M. G. Woldorff, "The spread of attention across modalities and space in a multisensory object," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18 751–18 756, December 20 2005.

[233] D. Soto and M. J. Blanco, "Spatial attention and object-based attention: a comparison within a single task," *Vision research*, vol. 44, no. 1, pp. 69–81, 2004.

[234] J. H. Maunsell and S. Treue, "Feature-based attention in visual cortex," *Trends in neurosciences*, vol. 29, no. 6, pp. 317–322, 2006.

[235] J. Theeuwes, "Feature-based attention: it is all bottom-up priming," *Philosophical transactions of the Royal Society of London.Series B, Biological sciences*, vol. 368, no. 1628, p. 20130055, September 09 2013.

[236] V. Maljkovic and K. Nakayama, "Priming of pop-out: I. role of features," *Memory & cognition*, vol. 22, no. 6, pp. 657–672, 1994.

[237] J. Theeuwes and E. Van der Burg, "The role of spatial and nonspatial information in visual selection," *Journal of experimental psychology.Human perception and performance*, vol. 33, no. 6, pp. 1335–1351, December 01 2007.

BIBLIOGRAPHY

[238] J. Theeuwes and E. Van der Burg, "On the limits of top-down control of vl selection," *Attention, perception & psychophysics*, vol. 73, no. 7, pp. 2092–2103, October 01 2011.

[239] A. V. Belopolsky, D. Schreij, and J. Theeuwes, "What is top-down about contingent capture?" *Attention, perception & psychophysics*, vol. 72, no. 2, pp. 326–341, February 01 2010.

[240] A. K. Lee, S. Rajaram, J. Xia, H. Bharadwaj, E. Larson, M. S. Hämäläinen, and B. G. Shinn-Cunningham, "Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch," *Frontiers in neuroscience*, vol. 6, 2012.

[241] E. Holmes, P. T. Kitterick, and A. Q. Summerfield, "Eeg activity evoked in preparation for multi-talker listening by adults and children," *Hearing research*, vol. 336, pp. 83–100, 2016.

[242] A. Clark, "Whatever next? predictive brains, situated agents, and the future of cognitive science," *Behavioral and Brain Sciences*, vol. 36, no. 03, pp. 181–204, 2013.

Vita

E. Merve Kaya received the Bachelors of Engineering in Computer Engineering from TOBB University of Economics and Technology in 2009, and Masters of Engineering in Electrical and Computer Engineering from Johns Hopkins University in 2011. She enrolled in the Electrical and Computer Engineering Ph.D. program at Johns Hopkins University in 2009, where she worked on machine learning and computer vision along with audio perception. Her research focuses on how humans attend to sound, and the application of biological insight into practical computational models.